РОСКОНГРЕСС
Пространство доверия

Большие языковые модели генеративного ИИ

От штучных изделий к стандартизированному товару

Февраль 2025

- Роль госфинансирования становится все важнее на фоне противостояния США и Китая в области искусственного интеллекта, хотя высокая активность венчурного капитала в этой сфере сохраняется.

- Санкции на поставку высокопроизводительных чипов в КНР простимулировали инновации, вынуждая разработчиков ИИ оптимизировать архитектуру моделей.

- Позиции американских разработчиков на рынке больших языковых моделей ослабевают на фоне ускоренного развития архитектуры со стороны китайских компаний.

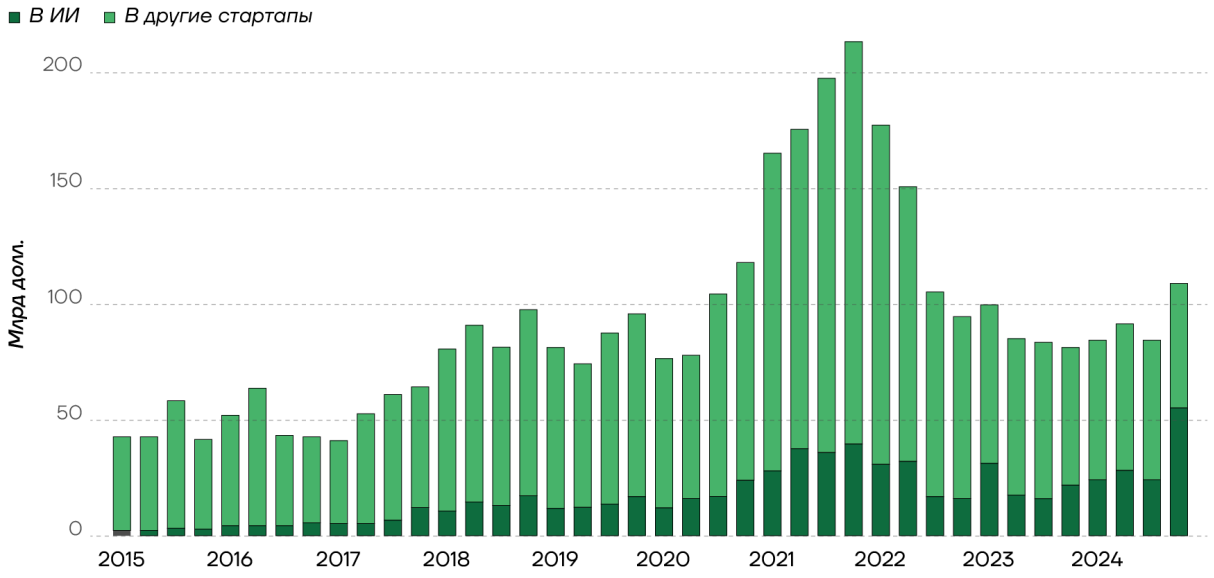
- Распространение открытых лицензий и удешевление использования больших языковых моделей переводят их из категории уникальных технологических продуктов в разряд стандартизированных инструментов.

Поток инвестиций не ослабевает

Сфера искусственного интеллекта в 2024 году сконцентрировала в себе львиную долю венчурного капитала. Согласно предварительным данным PitchBook, доля компаний, специализирующихся на ИИ, в общем объеме венчурных инвестиций в четвертом квартале 2024 года достигла 50,8%. Это почти вдвое превышает показатель 2023 года. При этом количество сделок в секторе ИИ снизилось на 16,6%, однако из-за сокращения совокупного объема венчурных инвестиций в мире их удельный вес увеличился не только в денежном, но и в количественном выражении — с 21,4% до 25,9%.

Объем венчурного финансирования стартапов в области ИИ достиг 131,5 млрд долларов США в 2024 году, продемонстрировав рост на 52% по сравнению с предыдущим годом. Данная динамика особенно показательна на фоне общего снижения венчурных инвестиций в другие направления примерно на 10% до 237 млрд долларов США. Основными получателями финансирования стали разработчики больших языковых моделей (large language model, LLM), включая Anthropic, Cohere и Mistral, а также компании, создающие инфраструктуру для работы с большими данными, такие как Databricks, привлекшая огромное финансирование на 10 млрд долларов США.

Квартальные объемы венчурного финансирования в мире



Источник: fDi intelligence, Pitchbook

LLM представляют собой систему из миллиардов вычислительных блоков или «нейронов», обученную на огромных массивах текстовых данных. Основой современных языковых моделей служит архитектура «трансформер», которая позволяет модели анализировать текст, учитывая взаимосвязи между словами в различных частях предложения или документа. Перед обработкой текст разбивается на небольшие фрагменты – токены, которые могут представлять собой как целые слова, так и их части или отдельные символы. В процессе работы модель использует механизм «внимания», позволяющий ей определять, какие части входного текста наиболее важны для формирования ответа. Размер текста, который модель может обрабатывать за один раз, ограничен так называемым контекстным окном – например, для некоторых современных моделей оно составляет до 200 тысяч токенов. При генерации ответа модель последовательно предсказывает наиболее вероятное продолжение текста на основе закономерностей, выявленных в процессе обучения. Подробнее про LLM можно узнать из исследования ИИМР «Генеративные нейросети: восстание машин или новая экономика.»

LLM способны оказать существенное влияние на экономику. По оценкам Goldman Sachs, внедрение таких моделей может привести к росту мирового ВВП на 7% в течение следующего десятилетия. Несмотря на высокие затраты на создание вычислительной инфраструктуры, экономический эффект от внедрения больших языковых моделей может быть значительным за счет

автоматизации рутинных задач, повышения производительности труда и трансформации бизнес-процессов. Ожидания серьезного экономического эффекта оправдывают существенные инвестиции в эту отрасль со стороны частного и государственного капитала.

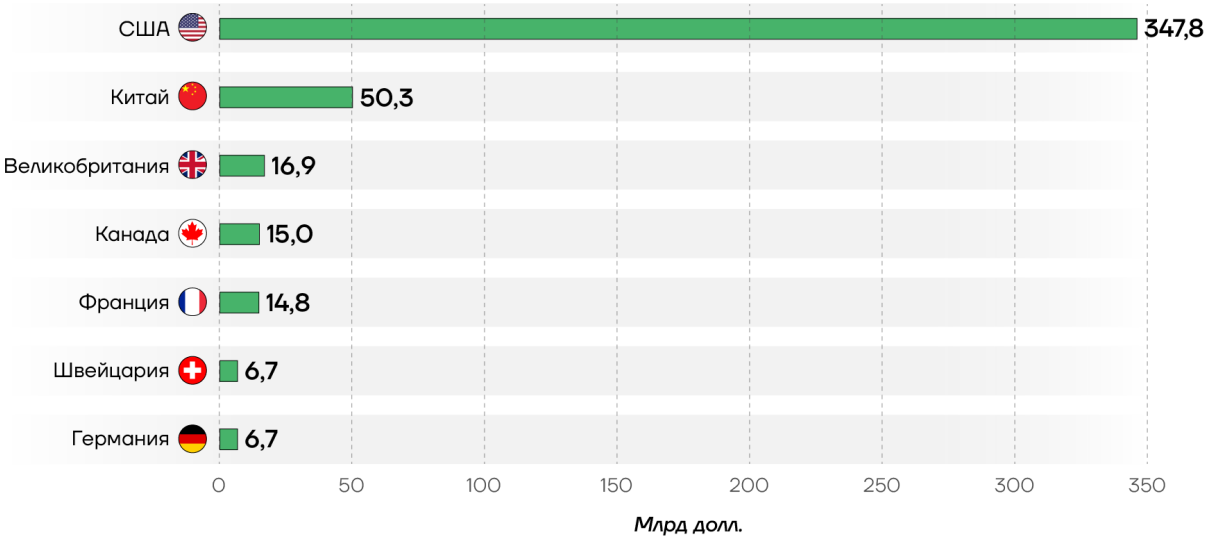
Примечательной особенностью инвестиционного ландшафта является активное участие технологических корпораций в финансировании ИИ-стартапов. Amazon, Microsoft, Google и Nvidia выступают в качестве стратегических инвесторов, что существенно отличает текущую ситуацию от традиционной модели венчурного финансирования. В частности, Nvidia в 2024 году инвестировала 1 млрд долларов США в 50 этапов финансирования стартапов.

Восточный и западный подходы к финансированию

Географическое распределение инвестиций характеризуется доминированием Северной Америки, где сконцентрировано около трети всех сделок и 60% объема венчурных инвестиций в ИИ-стартапы. Европа занимает второе место с долей около четверти от общего количества венчурных этапов в секторе ИИ. Высокие объемы финансирования обусловлены значительными затратами на создание инфраструктуры и сбор данных для обучения моделей, что создает существенные барьеры для входа в отрасль.

Распределение финансирования ИИ по странам

Январь-октябрь 2024 года

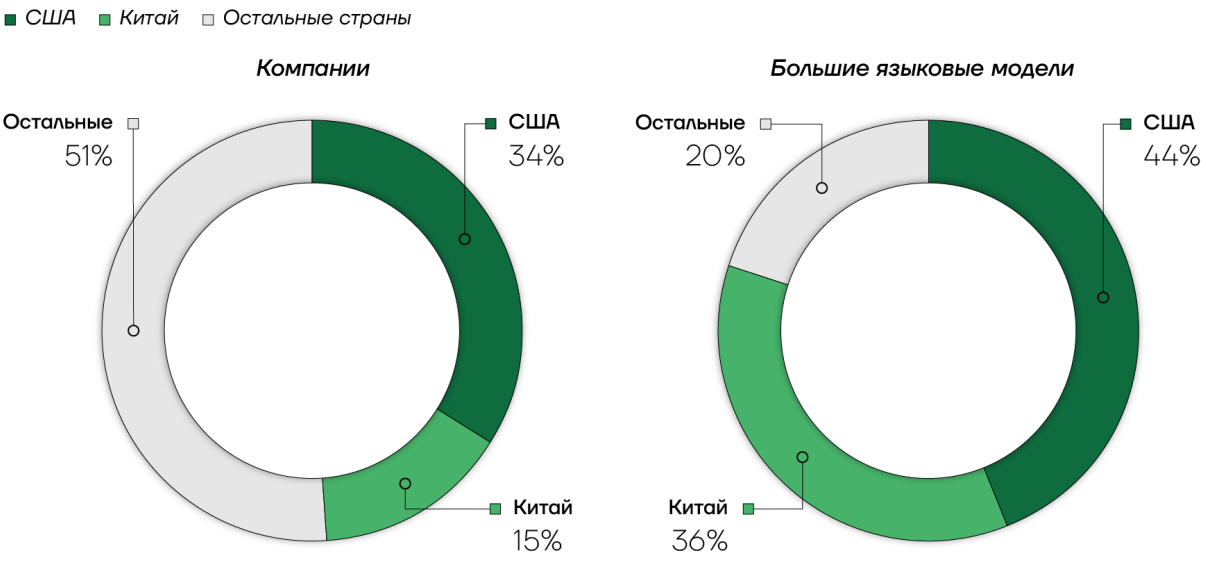


Источник: MoonFox Data

Финансирование через венчурные фонды не распространено в Китае, где также продолжается ИИ-бум. По общим вложениям в сектор КНР занимает второе место в мире, причем объем инвестиций больше, чем у трех крупнейших стран Европы, вместе взятых.

Принципиальное отличие китайской модели заключается в доминировании государственного и квазигосударственного финансирования через специализированные фонды развития и государственные банки. В отличие от американской модели, где ключевую роль играют частные венчурные фонды, китайские технологические компании получают основные средства через систему государственных институтов развития и крупных банков с госучастием. Китайская модель также характеризуется активным участием региональных властей, которые создают специальные зоны развития технологий и предоставляют значительные налоговые льготы.

Распределение компаний и моделей ИИ в мире



Источник: Китайская академия информационных и коммуникационных технологий (Белая книга по глобальной цифровой экономике 2024)

Эффективность такого подхода подтверждается статистическими данными: при меньшем количестве компаний в секторе искусственного интеллекта — 15% от мирового показателя против 34% у США — китайские разработчики создали сопоставимое количество языковых моделей — 36% против 44% у США. В совокупности на компании из этих двух стран приходится 80% таких моделей. Это свидетельствует о более высокой результативности государственно-ориентированной модели финансирования в контексте развития критически важных технологий.

Стратегии развития ИИ в США и КНР

В ближайшие годы к частным инвестициям в США будут добавлены государственные. Новая администрация Белого дома объявила о поддержке проекта Stargate. Он представляет собой совместное предприятие, основными участниками которого являются OpenAI, Oracle и SoftBank, при дополнительном участии эмиратского инвестиционного фонда MGX. Технологическими партнерами выступают Microsoft, Nvidia и Arm. Первоначальные инвестиции, предварительно, составят 100 млрд долларов с перспективой роста до 500 млрд долларов к 2029 году. Масштаб начинания подчеркивается планами строительства до 20 крупных дата-центров.

Среди ключевых рисков проекта следует выделить проблему привлечения заявленного объема финансирования, дефицит квалифицированных специалистов в области ИИ и потенциальные ограничения энергетической инфраструктуры. Примечательно, что федеральное правительство уже выразило готовность содействовать в решении энергетических вопросов, что может свидетельствовать о восприятии проекта как стратегически значимого для национальной безопасности.

Существенным фактором неопределенности остается бизнес-модель проекта. На данный момент не определены ни целевые потребители создаваемой инфраструктуры, ни конкретные продукты и услуги, которые будут предоставляться. Это затрудняет оценку потенциальной окупаемости инвестиций, которая, вероятно, станет возможной лишь в долгосрочной перспективе.

Китай реализует долгосрочную стратегию развития искусственного интеллекта, основанную на поэтапном подходе с четкими целевыми показателями. Ключевым документом является «План развития искусственного интеллекта нового поколения» от 2017 года, определяющий цели до 2030 года. К этому сроку планируется достичь мирового лидерства в сфере ИИ с объемом основной индустрии более 140 млрд долларов и связанных отраслей в 1,4 трлн долларов.

Китай создал масштабную инфраструктуру для развития ИИ, включающую 26% мировых вычислительных мощностей. Объем генерируемых данных демонстрирует среднегодовой рост в 26%, что является самым высоким показателем в мире.

При развертывании вычислительной инфраструктуры повышенное внимание уделяется энергоэффективности: ввод экологичных дата-центров, использование возобновляемых источников энергии, оптимизация систем охлаждения. Реализуется стратегическая инициатива «Восточные данные, западные вычисления» по перераспределению вычислительных ресурсов в регионы с доступной возобновляемой энергией.

Итогом китайской модели развития стало формирование сектора ИИ, который включает в себя около 4500 компаний. Среди них как стартапы в сфере искусственного интеллекта, так и гиганты, для которых ИИ является лишь частью бизнеса. Наиболее значимыми игроками являются Huawei, Baidu, DeepSeek, Moonshot AI, Zhipu AI, MiniMax, Baichuan AI и 01.AI.

Ведущей компанией сектора является Huawei, разрабатывающая модель PanGu, которая является мультимодальной, то есть она интегрирует в себе возможности обработки естественного языка, распознавания и генерации визуальных образов и звука.

Модель Ernie Bot (Wenxin Yiyan) от Baidu демонстрирует превосходные результаты в области генерации контента с использованием технологии извлечения информации (Retrieval Augmented Generation, RAG). Данная технология особенно эффективна в контексте поисковых систем и решения мультимодальных задач.

Особого внимания заслуживает компания DeepSeek, недавно представившая модели DeepSeek-V3 и DeepSeek-R1, которые по своим характеристикам составляют конкуренцию продуктам OpenAI.

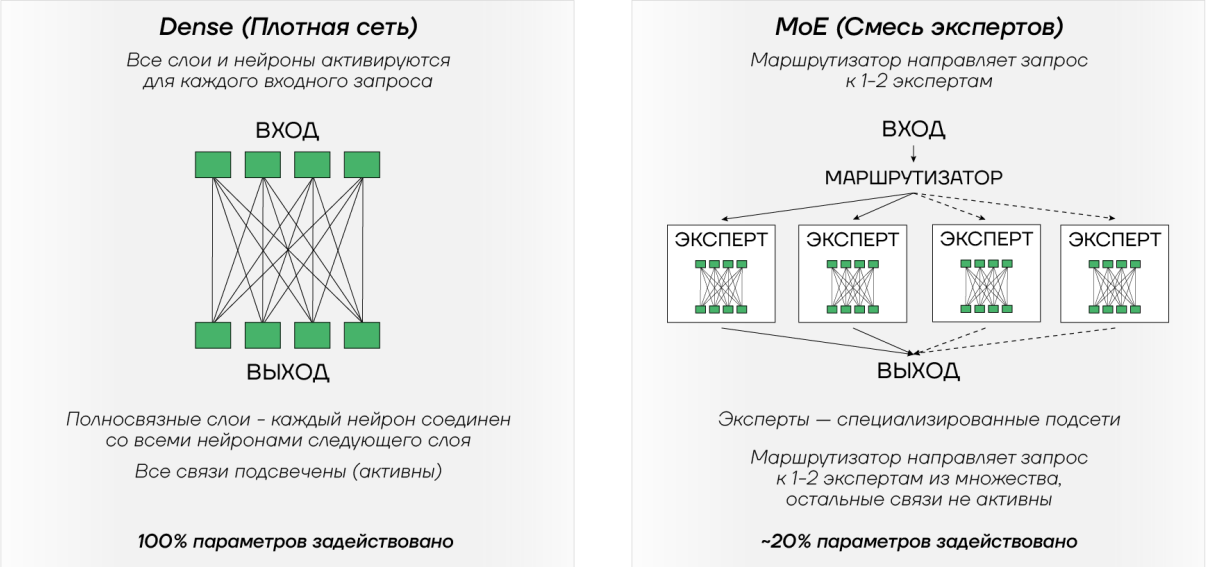
Значительный интерес представляет компания Moonshot AI, создавшая модель Kimi. Ключевым преимуществом данной разработки является способность обрабатывать объемные массивы текста в сочетании с возможностями мультимодального анализа.

Стоит отметить, что компании из КНР работают в условиях санкционного давления со стороны предыдущей и нынешней администраций Белого дома, и не могут напрямую импортировать самые мощные чипы. Внешние ограничения придают дополнительный импульс экспериментам над архитектурой и принципами работы ИИ-моделей.

Развитие эффективных архитектурных решений и методов обучения

В моделях DeepSeek-V3 и DeepSeek-R1 использован подход «смесь экспертов» (Mixture of Experts, MoE) вместо традиционной схемы с плотными связями (Dense).

Сравнение архитектур Dense и MoE



Источник: ИИМР

Последняя предполагает активацию всех параметров модели при каждом запросе, что обеспечивает стабильность и предсказуемость результатов. В то же время данный подход требует значительных вычислительных ресурсов и энергозатрат, что существенно влияет на экономическую эффективность их применения в промышленных масштабах.

Модели типа MoE используют динамическую маршрутизацию запросов к специализированным подсетям-экспертам, что позволяет существенно снизить вычислительную нагрузку при выполнении конкретных задач. При обработке запроса активируется лишь небольшая часть параметров модели, что теоретически должно приводить к снижению энергопотребления и стоимости вывода. На практике реализация подобных систем сопряжена с дополнительными сложностями в части распределения запросов между экспертами и обеспечения стабильности результатов.

Экономические параметры архитектур Dense и MoE

Dense (Плотная сеть)

Линейный рост вычислительных затрат с увеличением размера модели

Плюсы

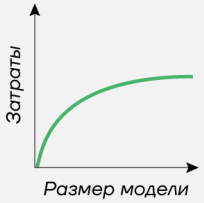
Предсказуемость, стабильность

Минусы

Высокие затраты на вычисления

MoE (Смесь экспертов)

Возможность добавления экспертов без пропорционального роста затрат на вывод



Плюсы

Экономия на выводе за счет частичной активации

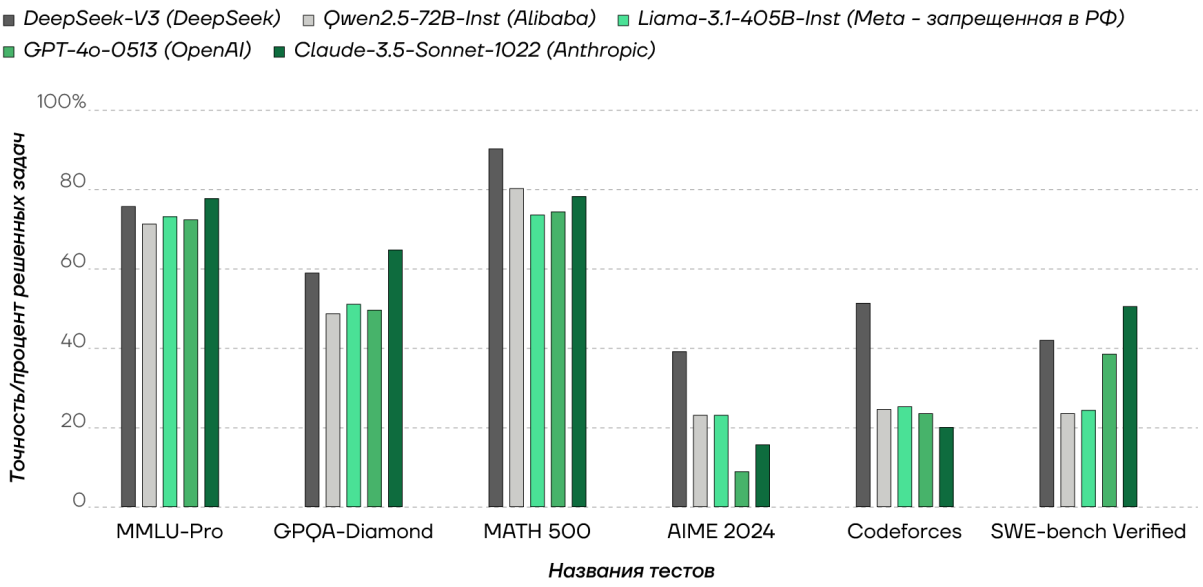
Минусы

Сложность балансировки нагрузки между экспертами
Деградация качества при ошибках маршрутизации

Источник: ИИМР

Существенным преимуществом MoE является возможность масштабирования модели без пропорционального роста вычислительных затрат при выводе. При этом данный подход требует более сложной инфраструктуры для эффективного распределения нагрузки между экспертами и координации их работы. Дополнительным фактором риска выступает потенциальная деградация качества при неоптимальной маршрутизации запросов или перегрузке отдельных экспертов. Подход MoE уже использовала компания Google в моделях GShard или Switch Transformer, однако DeepSeek реализовала его удачнее американских конкурентов и смогла добиться относительно стабильной работы системы.

Результаты тестирования LLM



Источник: DeepSeek

Судя по результатам самых распространенных тестов DeepSeek-V3 демонстрирует приблизительно такие же результаты, что и другие современные модели. В тесте MATH 500, где нужно решать математические задачи уровня старшей школы, точность достигает 90,2%, что существенно превышает результаты других систем искусственного интеллекта.

Стоит отметить относительно слабую разницу результатов между моделями Qwen2.5-72B-Inst, Llama-3.1-405B-Inst и GPT-4o-0513 в большинстве тестовых категорий, что может свидетельствовать о достижении определенного технологического плато в рамках используемых архитектурных решений. Однако модель Claude-3.5-Sonnet-1022 демонстрирует нетипичную динамику, показывая значительное превосходство в тесте на решение задач для программистов SWE-bench Verified с результатом 50,8%.

Появление DeepSeekV3 на архитектуре MoE задало новый уровень цен для компаний, которые предоставляют доступ к LLM. Использование самых современных моделей от OpenAI обходится на порядок или два дороже, чем аналогичных по качеству моделей от DeepSeek. Кроме того, китайская компания опубликовала свои модели в свободном доступе под лицензией MIT, которая разрешает их бесплатное использование и модификацию в любых целях, включая коммерческие. Производные модели на основе разработок от китайской компании можно лицензировать на любых условиях.

Стоимость пользования LLM

Провайдер	Модель	Контекстное окно	Ввод млн токенов	Вывод млн токенов
OpenAI	o1	200K/100K	$15	$60
OpenAI	o1-mini	128K/65K	$3	$12
Anthropic	Claude 3.5 Sonnet	200K/8K	$3	$15
Anthropic	Claude 3.5 Haiku	200K/8K	$0.8	$4
Meta via Deepinfra	Llama 3.1 405b	200K/8K	$1.79	$1.79
DeepSeek	DeepSeek-R1	128K/2K	$0.55	$2.19
DeepSeek	DeepSeek-V3	128K/8K	$0.14	$0.28

Источник: docsbot.ai, дата обращения 06.02.2025

Таким образом бизнес-модель OpenAI и других провайдеров ИИ ставится под угрозу. Дополнительным фактором, задающим тренд на удешевление пользования LLM, является выпуск так называемых «дистиллированных» моделей, которые используют меньшее число параметров, сохраняя при этом большую часть полезных свойств.

Ключевым механизмом процесса является передача знаний от «учительской» модели к «студенческой» посредством имитации выходных данных крупной модели меньшей моделью. Самым распространенным подходом является метод «дистилляции знаний» (Knowledge Distillation, KD). Принципиальная особенность данного метода заключается в использовании двойственного подхода к обучению «студенческой» модели: одновременно применяются как вероятностные распределения выходных данных «учительской» модели (мягкие цели), так и фактические правильные ответы из обучающего набора данных (жесткие цели).

Экономическая ценность такого подхода обусловлена тем, что мягкие цели содержат более богатую информацию о процессе принятия решений исходной моделью, включая относительную уверенность модели в различных вариантах ответа. Это позволяет «студенческой» модели усваивать не только конечные результаты, но и нюансы функционирования «учительской» модели, что принципиально повышает качество дистилляции при сохранении эффекта сокращения вычислительных затрат.

Схема «дистилляции» LLM

Источник: datacamp

Использование вероятностных распределений вместо единичных правильных ответов в процессе обучения обеспечивает более плавный процесс передачи знаний и, как следствие, более надежную работу результирующей уменьшенной модели. В контексте промышленного применения это транслируется в повышенную надёжность и предсказуемость работы систем на основе дистиллированных моделей.

Экономическая значимость данной технологии уже была продемонстрирована DeepSeek, чья модель R1, созданная с применением дистилляции, показала сопоставимые с OpenAI o1 результаты при радикально меньших затратах. Стоимость разработки R1 оценивается в 6 млн долларов против 500 млн долларов на создание o1. Операционные расходы также существенно ниже: стоимость обработки миллиона токенов составляет 2,19 доллара для R1 против 60 долларов для o1.

Применение дистилляции позволяет значительно оптимизировать использование вычислительных ресурсов. При сохранении ключевых возможностей уменьшенная модель требует меньше памяти для хранения и обеспечивает более быстрый отклик. Это открывает возможности для внедрения языковых моделей на устройствах с ограниченными ресурсами, таких как мобильные телефоны и краевые вычислительные устройства.

Существенным фактором является возможность создания дистиллированной модели без доступа к исходным обучающим данным. Это создаёт потенциал для появления новых игроков на рынке языковых моделей, способных конкурировать с крупными технологическими компаниями при меньших инвестициях. По оценкам специалистов Microsoft Research, это может привести к коммодитизации крупных языковых моделей, смещая фокус создания экономической ценности в сторону разработки приложений на их основе.

Интеллектуальная собственность в эпоху ИИ

Отдельного внимания заслуживает правовой аспект: создание дистиллированных версий существующих моделей поднимает вопросы интеллектуальной собственности, которые могут существенно повлиять на развитие рынка искусственного интеллекта и распределение экономических выгод между участниками. В конце января генеральный директор OpenAI Сэм Альтман заявил

о возможном неправомерном использовании компанией DeepSeek результатов работы моделей OpenAI для создания собственных продуктов. В начале февраля он рассказал журналистам, что не планирует подавать в суд на китайскую компанию.

В более широком контексте данная ситуация отражает общую проблематику прав интеллектуальной собственности в сфере искусственного интеллекта. Аналогичные вопросы возникают в отношении использования общедоступных данных для обучения моделей, что иллюстрируется судебными исками New York Times и других издателей к OpenAI.

Ожидается принятие американскими технологическими компаниями мер по предотвращению несанкционированной дистилляции моделей. Это может существенно замедлить появление конкурирующих продуктов на рынках США и союзников. В контексте международной конкуренции данный вопрос приобретает стратегическое значение, что подтверждается вовлечением правительственных структур США в защиту технологических преимуществ американских компаний.

Сенаторы-республиканцы в конце января внесли на обсуждение Конгресса США законопроект S.321 «О разделении возможностей искусственного интеллекта Америки и Китая» (Decoupling America's Artificial Intelligence Capabilities from China Act). Он предусматривает три ключевых направления ограничений. Во-первых, вводится полный запрет на импорт технологий и интеллектуальной собственности в сфере ИИ, разработанных или произведенных в Китае, а также на экспорт аналогичных американских разработок в КНР. Во-вторых, запрещается проведение исследований и разработок в области ИИ на территории Китая или в сотрудничестве с китайскими организациями. В-третьих, вводится запрет на финансирование американскими лицами китайских разработок в сфере ИИ через инвестиции, кредиты или управление активами.

Особенностью законопроекта является его широкая юрисдикция — под определение «американского лица» подпадают не только граждане и резиденты США, но также корпорации, образовательные и исследовательские учреждения, зарегистрированные в США или контролируемые американскими гражданами. За нарушение запретов предусмотрены штрафы до 100 млн долларов для организаций и до 1 млн долларов для физических

лиц, а также лишение государственных контрактов, грантов и других форм федеральной поддержки на 5 лет.

Экономические последствия принятия данного закона могут быть значительными, учитывая масштаб существующего технологического сотрудничества между странами. Законопроект фактически требует полного разрыва связей в сфере ИИ, что затронет как крупные технологические компании, так и научно-исследовательское сообщество.

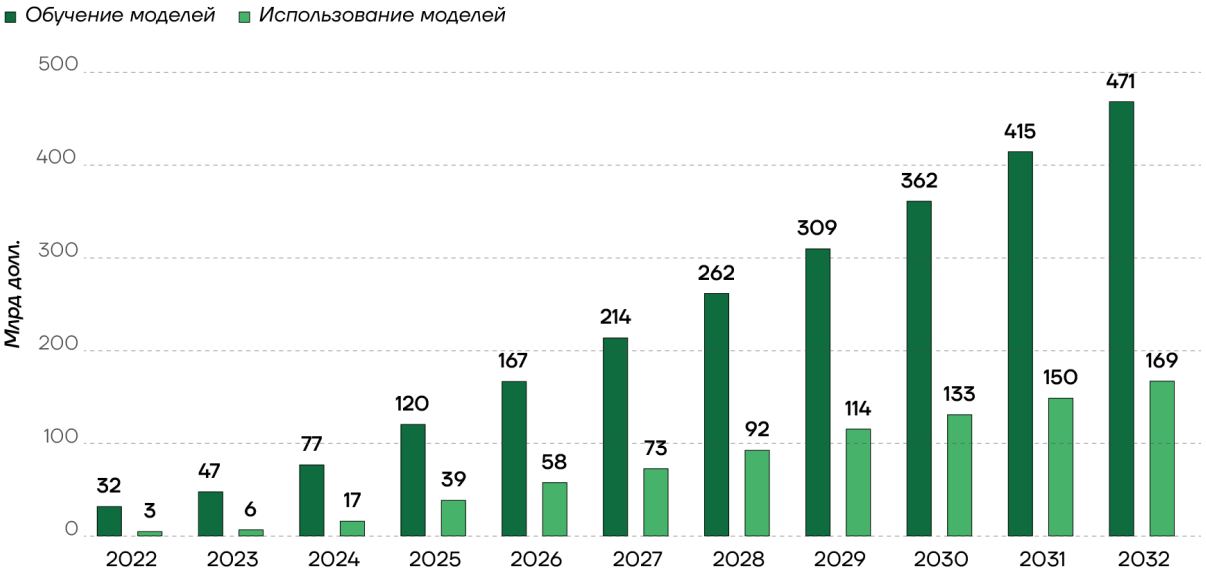
Хотя перспективы принятия законопроекта в текущем виде неясны, он отражает растущее стремление Конгресса США к технологическому разделению с Китаем. Об этом свидетельствуют также недавние обращения конгрессменов к администрации с призывами усилить экспортный контроль и запретить использование китайских ИИ-систем в государственных закупках.

Радужные прогнозы сектора ИИ от западных аналитиков

Рынок генеративного искусственного интеллекта, по прогнозам аналитиков Bloomberg, к 2032 году достигнет 1,3 трлн долларов США, что составит 10-12% всех технологических расходов. Наиболее существенную долю этого рынка, около 471 млрд долларов, займет сегмент обучения LLM. Второй по значимости составляющей станет сегмент инфраструктуры как услуги для генеративного ИИ, который достигнет 309 млрд долларов при среднегодовом темпе роста в 54%. Стоит отметить, что даже американские аналитики не ожидают, что ключевым источником выручки станет непосредственно использование генеративного ИИ.

Экономическая модель OpenAI и других провайдеров LLM принципиально отличается от традиционных программных продуктов. В то время как классические технологические компании при масштабировании бизнеса демонстрируют снижение маржинальных издержек, у компаний сектора ИИ наблюдается синхронный, а в некоторых случаях опережающий рост затрат по отношению к выручке.

Выручка на рынке генеративного ИИ



Источник: Bloomberg Intelligence

Показательным примером является OpenAI, которая при прогнозируемой выручке в 3,7 млрд долларов в 2024 году ожидает убытки в размере 5 млрд долларов, причем только затраты на вычислительные мощности составят 6 млрд долларов. Как стало известно во время этапа финансирования, завершившегося в октябре 2024 года, на каждый доллар выручки компания тратит 2,35 доллара. Затраты на обучение модели GPT-4 составили 100 млн долларов, при этом прогнозировался рост расходов на обучение моделей до 3 млрд долларов в прошедшем году.

Заявленные компанией прогнозы роста выручки до 11,6 млрд долларов в 2025 году и 100 млрд долларов к 2029 году представляются необоснованными с учетом текущей структуры затрат. Даже при трехкратном увеличении выручки к концу 2025 года и оптимистичном сценарии двукратного снижения издержек, убыток компании составит не менее 2 млрд долларов. При этом рост базы пользователей бесплатной версии ChatGPT создаст дополнительную нагрузку на расходную часть.

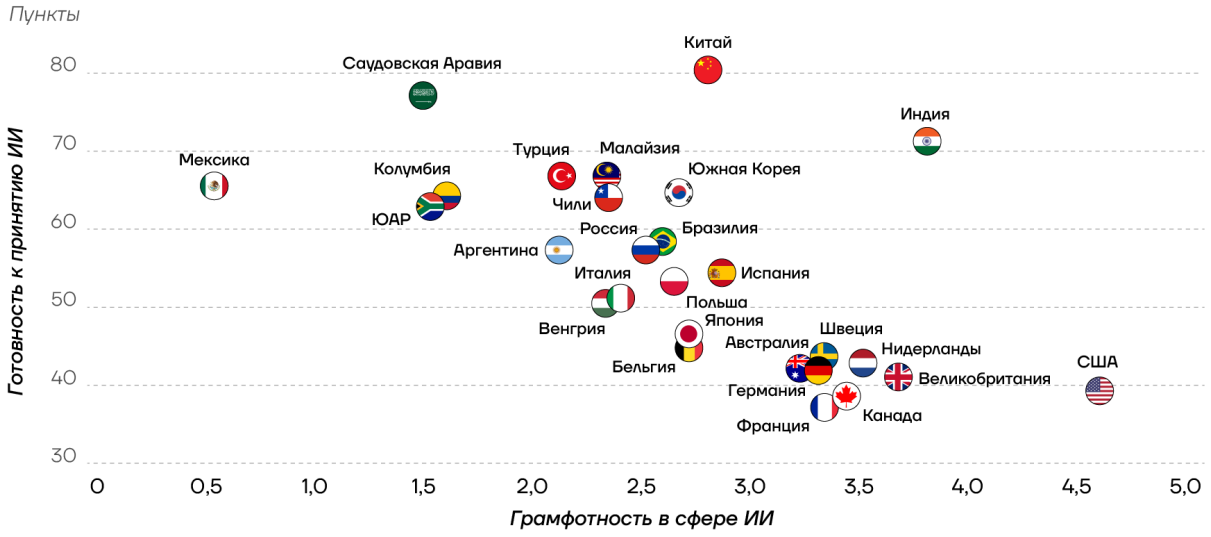
Особую обеспокоенность вызывает структура выручки OpenAI. Изначально основным способом монетизации должен был стать платный доступ к LLM с помощью API (application programming interface), технологии, позволяющей интегрировать возможности модели в стороннее приложение. Однако через этот канал компания получила менее 1 млрд долларов или около 30% общей выручки. Данный показатель может свидетельствовать о системных ограничениях роста индустрии генеративного искусственного интеллекта

в целом. Планируемое повышение цен доступа через API может сделать их экономически нецелесообразными для значительного числа компаний, в настоящее время интегрирующих технологии OpenAI в свои продукты. Клиенты могут либо запускать «дистиллированные» модели на собственном оборудовании, либо использовать одну из моделей с «открытой» лицензией на арендованных в дата-центрах мощностях.

Неграмотность как драйвер роста

Рост капитализации компаний в секторе во многом основан на фантастических надеждах, связанных с завышенными ожиданиями от LLM. Доценты американских бизнес-школ в исследовании «Более низкая грамотность в области искусственного интеллекта предсказывает большую восприимчивость к нему» (Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity), основанном на изучении данных опросов Ipsos, выявили парадоксальную закономерность: чем ниже уровень осведомленности о технологических составляющих в сфере ИИ, тем больше готовность к его принятию и использованию. Данный феномен объясняется тем, что люди с меньшим пониманием принципов работы ИИ склонны воспринимать его как нечто магическое, вызывающее трепет и восхищение.

Низкая грамотность в сфере ИИ способствует его принятию



Источник: Tully, S., Longoni, C., & Appel, G. (2023, June 16). Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity

Исследование демонстрирует, что традиционный подход к продвижению ИИ через объяснение его преимуществ и принципов работы может быть

контрпродуктивным. Демистификация технологии снижает её привлекательность для потенциальных пользователей и инвесторов. Это подтверждается на примере британского плана действий в области ИИ, который случайно или преднамеренно описывает языковые модели как магические сущности, которые имеют собственные желания, эмоции, думают и говорят.

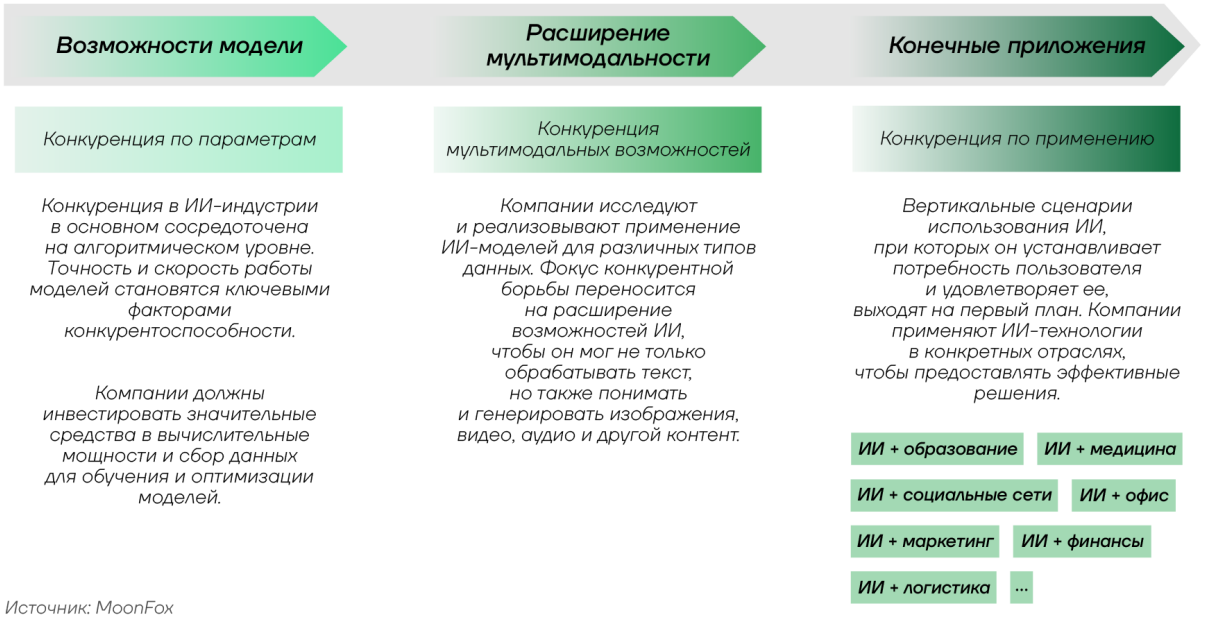
Ключевым фактором успешного продвижения и маркетинга ИИ становится антропоморфизация — наделение систем искусственного интеллекта человеческими чертами и создание иллюзии личности. Этот подход особенно эффективен для аудитории с низким уровнем технической грамотности, которая стремится к эмоциональному взаимодействию с технологией.

Данные выводы имеют серьезные последствия для стратегии маркетинга ИИ-продуктов. Вместо образовательного подхода и разъяснения технических возможностей, более эффективным может быть создание впечатляющих демонстраций, вызывающих эмоциональный отклик и чувство благоговения. При этом следует избегать детальных объяснений принципов работы технологии, поскольку это может разрушить ореол «магии» вокруг ИИ.

От параметров к применению

В аналитическом обзоре индустрии искусственного интеллекта, совместно подготовленном агентством MoonFox и Центром исследования ИИ и управленческих инноваций Китайско-европейской международной бизнес-школы (China Europe International Business School, CEIBS), указано, что в Китае акцент смещается от конкуренции платформ к конкуренции приложений. Другими словами, происходит переход от создания и оптимизации базовых моделей к проектированию и продвижению конечных пользовательских приложений. Эта трансформация демонстрирует рыночный спрос на практические результаты применения ИИ. Компании и платформы больше не фокусируются исключительно на параметрах и вычислительных возможностях моделей, а уделяют больше внимания тому, как превратить эти технологии в приложения, которые приносят ощутимый коммерческий эффект и удовлетворяют запросы пользователей.

Эволюция конкуренции в ИИ-индустрии КНР

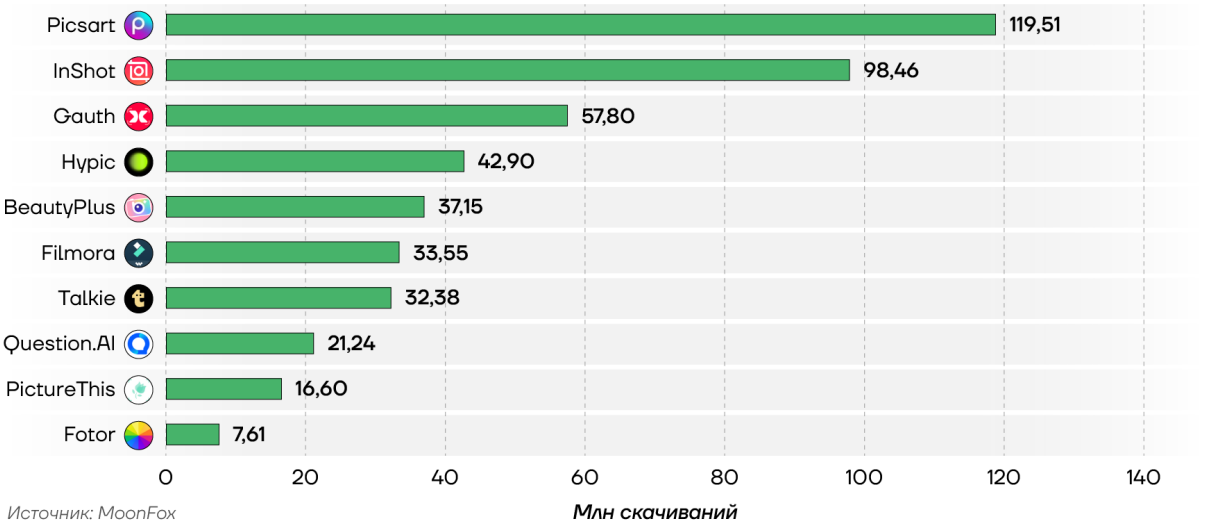


Источник: MoonFox

Создав востребованные продукты для внутреннего потребления, китайские разработчики ИИ-приложений уже выходят на мировой рынок. Они могут повторить успех TikTok от ByteDance, который стал одной из крупнейших социальных сетей в мире благодаря эффективному алгоритму рекомендации контента и столкнулся с противодействием американских властей. С 19 января де-юре это приложение находится в США под запретом из-за опасений потенциального сбора пользовательских данных. 20 января, в первый день своего президенства, Дональд Трамп подписал указ, который приостановил действие запрета на 75 дней.

Количество скачиваний китайских ИИ-приложений за пределами КНР

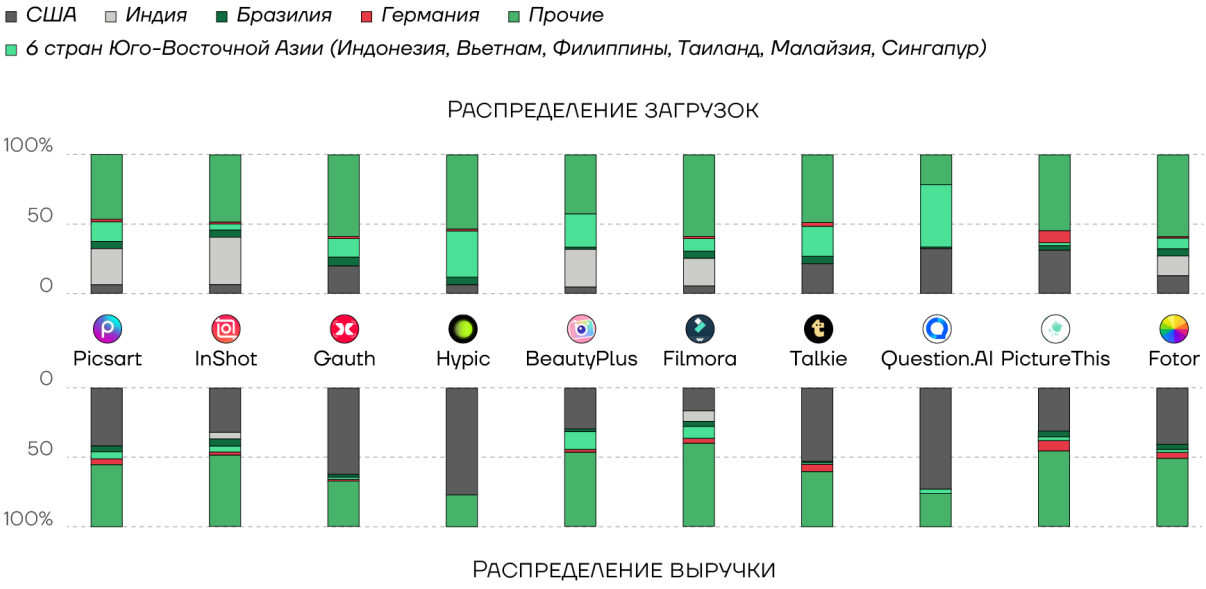
Январь-ноябрь 2024 года



Источник: MoonFox

По состоянию на второе полугодие 2024 года количество посещений китайских ИИ-приложений через веб-сайты стабильно превышает 180 миллионов в месяц, с пиковыми значениями более 200 миллионов в сентябре и октябре. Мобильные приложения демонстрируют устойчивый рост загрузок, что указывает на значительный потенциал данного канала дистрибуции. Лидерами по количеству установок являются утилиты Picsart и InShot, а также образовательное приложение Gauth.

Распределение загрузок и выручки в приложениях



Источник: MoonFox

США является основным рынком, приносящим доход китайским ИИ-приложениям, в Южной и Юго-Восточной Азии еще не сформировалась привычка платить за программы для смартфонов, хотя именно на эти страны приходится большинство загрузок.

Можно предположить, что борьба между ИИ-компаниями из США и КНР за кошельки американцев в ближайшее время продолжит обостряться, и в ходе конкурентной борьбы администрация Белого дома будет применять административные меры.

Выводы

- Развитие архитектуры больших языковых моделей и методов дистилляции знаний существенно снижают барьеры входа в эту сферу, способствуя переходу от рынка, на котором доминирует крайне малое число компаний, к более конкурентной среде.

- Экономическая модель действующих провайдеров LLM демонстрирует проблемы масштабирования: рост выручки сопровождается пропорциональным или опережающим ростом затрат, что ставит под сомнение долгосрочную устойчивость их текущей бизнес-модели.

- Противостояние между США и Китаем в сфере ИИ приобретает институциональные формы через законодательные инициативы и механизмы защиты интеллектуальной собственности, что может привести к формированию двух параллельных технологических экосистем.

- Наблюдается дивергенция между потребительским и корпоративным сегментами рынка LLM: если в потребительском сегменте доминирует тренд на упрощение и удешевление доступа к технологии, то в корпоративном растет спрос на специализированные решения с высокой степенью контроля над данными и процессами.

- Успешному продвижению технологий искусственного интеллекта способствует низкий уровень технической грамотности потребителей и инвесторов, воспринимающих LLM как «магическую» технологию, что определяет особенности маркетинговых стратегий в данном секторе.