Современная электроника и искусственный интеллект
Часть 1. Что такое искусственный интеллект, и что он может

Виктор Алексеев

Проявления искусственного интеллекта (ИИ) мы замечаем всё чаще как в повседневной жизни, так и в самых различных областях науки, техники, медицины, транспорта и т.д. Общая цель нескольких частей этой статьи заключается в том, чтобы попытаться объяснить, с одной стороны, какую роль играет современная электроника в ИИ, а с другой – как развитие современной электроники влияет на прогресс ИИ.
В первой части статьи простыми словами рассказано, что такое ИИ, и как он работает. На основе анализа статей ведущих экспертов в области искусственного интеллекта автор постарался выделить несколько наиболее крупных фирм, продукция которых представляется наиболее перспективной. В первой части приведён краткий обзор больших языковых моделей (LLM) этих фирм. В следующей части статьи планируется рассмотреть специализированные модели искусственного интеллекта.

Современный искусственный интеллект

В последнее время наблюдается стремительный рост различных приложений на базе так называемого искусственного интеллекта – ИИ (Artificial Intelligence – AI), сопровождаемый многочисленными дискуссиями на тему «хорошо это или плохо».

Для того чтобы ответить на этот вопрос, прежде всего нужно постараться определить, что означает сам термин «Artificial Intelligence».

Считается, что термин «Artificial Intelligence» впервые был официально введён Джоном Маккарти в 1956 году на конференции в Дартмутском колледже, хотя концептуальные основы этого научного направления были заложены ранее, в первую очередь благодаря работам Алана Тьюринга [1].

В те годы термин «Artificial Intelligence – AI» однозначно определял конкретные инструменты и программы, облегчающие решение определённых задач, например, компьютер на транзисторах (NCR-304, 1957).

С течением времени AI постепенно развивался, позволяя решать всё более сложные задачи: от чат-ботов, отвечающих на простые вопросы, до мощных систем, управляющих производством (рис. 1).

В переводе на русский английский термин «Intelligence» имеет множество различных значений, среди которых можно отметить следующие: разум; нечто, обладающее разумом; умственные способности; логико-информационные возможности; способность анализировать предоставленную информацию. Обобщая эти значения, коротко можно сказать, что «Intelligence» ближе всего к такому базовому понятию, как способность приобретать новые знания и навыки на основе накопленного опыта.

Для русского термина «интеллект» в английском языке одним из вариантов перевода является многозначный термин «Intellect», обобщающий понятие способности к мышлению, рассуждению и объективному пониманию окружающей действительности. Одним из свойств, которыми обладает интеллект человека (Human Intellect), является абстрактное мышление, характеризующее когнитивные способности именно человека.

По существу, термин «Intelligence» – это широкое понятие, охватывающее множество эмоциональных и практических аспектов, таких как, например, способность учиться на опыте, адаптация к новым ситуациям, способность решать возникающие проблемы, понимание и обработка абстрактных концепций, эмоциональная и социальная осведомлённость, распознавание образов и память.

Поэтому Intelligence развивается через различный собственный опыт и самостоятельное взаимодействие с внешним миром, а Intellect, в первую очередь, развивается посредством формального образования и обучения.

Несмотря на явные различия, перевод «Artificial Intelligence» как «искусственный интеллект» прижился в российских публикациях. В дальнейшем в этой статье мы будем пользоваться аббревиатурой «ИИ».

Эти и другие технические термины подробно описаны в документе «Национальная стратегия развития ИИ на период до 2030 года» [3].

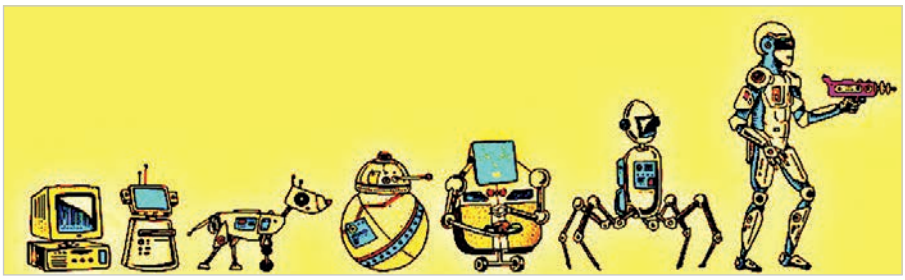

Рис. 1. Эволюция ИИ от первого компьютера до промышленных роботов [2]

Рис. 2. Современный ИИ – это разум или интеллект? [14]

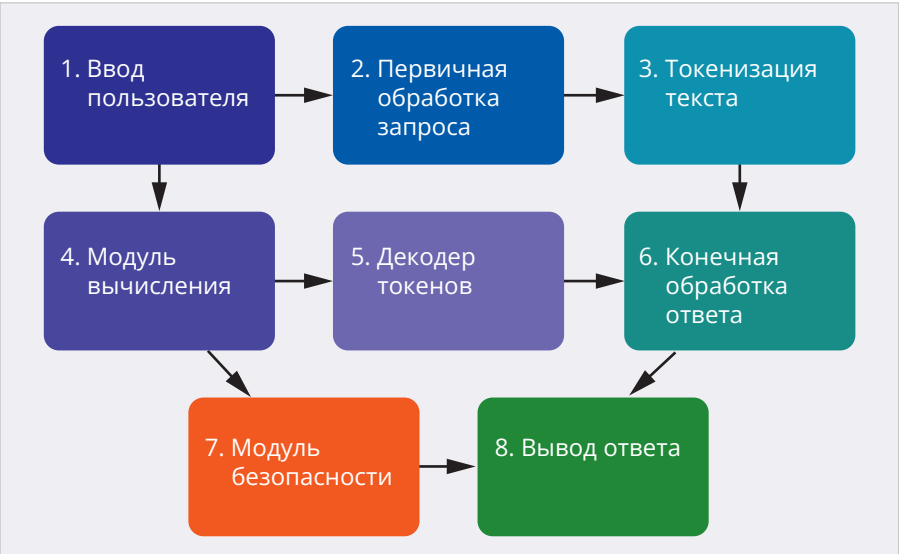

Рис. 3. Основные этапы работы современного искусственного интеллекта

Сегодня ИИ активно развивается не только за рубежом, но и в Российской Федерации, где этому направлению уделяется огромное внимание. На последней (11 декабря 2024) международной конференции «Путешествие в мир искусственного интеллекта», на которой присутствовали члены правительства, были продемонстрированы платформы искусственного интеллекта, разработанные российскими специалистами [4].

На выставке, которая прошла в рамках конференции, были представлены следующие российские AI: GigaChat, СБЕРБАНК [5]; Kandinsky 3.1 (Сбербанк) [6]; Yandex GPT-2 [7]; «Телематика» ComNews [8]; AI.T-bank – Центр искусственного интеллекта Т-Банка [9].

Российские разработки ИИ достаточно тщательно рассмотрены в многочисленных публикациях [10, 11].

Независимое тестирование, проведённое компанией «Зерокодер», показало, что российские LLM модели GigaChat и Yandex GPT в плане простых задач работают не хуже таких известных мировых AI, как ChatGPT-o1 и Llama 3.2. При этом стоимость платных версий российских продуктов значительно ниже, а доступность на территории РФ не ограничена [12].

Особую роль в развитии российских средств искусственного интеллекта сыграл созданный пять лет назад «Альянс в сфере искусственного интеллекта», в который, например, входят: Сбербанк; Т-Банк; Яндекс; МТС; Газпром; СИБУР; ВК; МТС; УРАЛХИМ, РУСАГРО; Северсталь; Ростелеком, Касперский и другие.

Необходимо отметить, что координацией деятельности по развитию ИИ занимается также автономная некоммерческая организация «Цифровая экономика», уполномоченная Правительством РФ.

По итогам конференции «Путешествие в мир искусственного интеллекта» (11.12.2024) был разработан ряд поручений президента правительству, направленных на развитие искусственного интеллекта в Российской Федерации [13].

В последние годы ИИ находит применение в самых различных новых областях науки, техники, медицины, транспорта, космоса, вооружений и т.д.

Сегодня AI умеет поддерживать диалог с человеком в реальном масштабе времени, обсуждая практически любые вопросы.

Возникает вопрос, что же такое современный Artificial Intelligence? Это по-прежнему только набор инструментов и программ, облегчающий труд человека? Или это уже некий новый феномен в развитии технологий, способный понимать контекст, учиться на опыте, адаптироваться к неизвестным ранее явлениям? (рис. 2) [14]

Ответ на этот вопрос является крайне важным с точки зрения ожиданий и разочарований как пользователей, так и разработчиков AI.

Что такое искусственный интеллект с технической точки зрения

От первого компьютера NCR-304 (1957), который выполнял лишь базовые математические операции, современный ИИ отличается масштабом и сложностью, но не своей фундаментальной природой: это по-прежнему программно-аппаратный комплекс, работающий по заданиям человека.

Эти системы позволяют обрабатывать естественный язык и генерировать осмысленные ответы, имитируя человеческое общение.

Основные этапы работы современного ИИ показаны на рис. 3, сгенерированном AI Claude 3.7.

Процесс начинается с ввода пользователя, формулирующего запрос через соответствующий интерфейс (веб-приложение, мобильное приложение, API и т.д.). Запрос может содержать текст, а в некоторых моделях также изображения или другие типы данных.

На этапе первичной обработки запроса происходит предварительная подготовка пользовательского запроса, включающая такие технические операции, как, например: нормализация текста (приведение к единому регистру, удаление лишних пробелов); предварительный анализ содержимого на предмет нарушения правил использования; добавление системных инструкций, исходя из имеющегося контекста; предварительное отсеивание запрещённого контента.

Важным этапом является токенизация текста, в процессе которой текст разбивается на токены, определяемые как минимальные текстовые единицы, с которыми работает модель. В качестве аналогии токенов можно привести разбиение предложения на части подобно тому, как мы разбиваем числа в математике. Токенами могут быть слова, части слов или отдельные символы, в зависимости от используемого «токенизатора». Например, английскую фразу «The cat sits on the mat» искусственный интеллект Claude разбивает на следующий набор токенов: «The», «cat», «sit», «s», «on», «the», «mat». В этом блоке также реализуется обработка специальных токенов, определяются начало и конец предложения. В отдельную группу выделяются неизвестные модели токены.

Каждая модель ИИ использует определённый набор специализированных алгоритмов токенизации, таких как, например, BPE, WordPiece, SentencePiece и другие.

На следующем этапе происходит подготовка информации к тензорным вычислениям, в результате которой токены преобразуются в многомер-

Рис. 4. Структурная схема аппаратной и программной частей современного ИИ

ные векторы (эмбеддинги). Не вдаваясь в подробности тензорной алгебры, можно отметить, что, например, наша токенизированная «cat» превратится в многомерный вектор [0,2; –0,5; 0,8; ...].

Основной этап обработки реализуется в аппаратном модуле вычисления, где происходит обработка запроса с использованием нейронной сети модели искусственного интеллекта. В результате модельных расчётов в декодер токенов поступает информация, позволяющая генерировать выходные токены.

Сгенерированный ответ модели проходит дополнительную обработку, необходимую для улучшения качества и безопасности в блоке «конечной обработки». На этом этапе осуществляются следующие технические операции: форматирование и структурирование ответа; проверка согласованности и логичности; контроль правил стилистики и грамматики; оформление ответа в соответствии с требуемым форматом.

Дополнительный контроль содержимого ответа в плане обеспечения безопасности и соответствия этическим нормам реализуется в модуле безопасности.

Вывод ответа на запрос клиента осуществляется на финальном этапе работы в форме, преобразованной в формат, понятный клиентскому приложению. Для реализации этих сложных процессов ИИ требует мощной аппаратной базы, о которой пойдёт речь далее.

Понимание архитектуры и принципов работы такой системы необходимо для эффективного проектирования, разработки и использования современных интеллектуальных систем на основе больших языковых моделей.

Структурная схема аппаратной и программной части современного ИИ показана на рис. 4, сгенерированном AI Claude 3.7.

Типичная аппаратная конфигурация современного ИИ (LLM) включает серверную часть, инфраструктуру и вспомогательное оборудование. Каждая из многочисленных фирм, занимающихся разработками ИИ, использует свои варианты компоновки аппаратурной части. Поэтому в этом описании мы ограничимся только общими хорошо известными схемами.

В современных больших языковых моделях используются графические процессоры (GPU), предназначенные для параллельной обработки с использованием матричных вычислений.

В качестве примера можно привести два наиболее мощных на сегодняшний день графических процессора NVIDIA H100 и H200.

Модель H100 с ядрами Tensor четвёртого поколения, созданная по технологии TSMC 4N, имеет 80 ГБ памяти HBM3 с пропускной способностью 3,35 Тбайт/с. Более новая версия H200 отличается увеличенным объёмом памяти HBM3e (до 141 ГБ), повышенной пропускной способностью памяти (до 4,8 Тбайт/с), а также более высокой энергоэффективностью. Обе модели имеют одинаковый конструктив «8U Rackmount». Кроме уникальных технических характеристик H100/H200 отличаются и крайне высокой стоимостью.

Конструкция NVIDIA H100 предусматривает совместную работу нескольких GPU, объединённых в стандартные серверные кейсы. Так, новая модель сервера NVIDIA DGX H100/H200 содержит восемь H100 или H200 графических процессоров (рис. 5). Этот сервер имеет габаритные размеры 897×356×482 мм и весит 131 кг.

Управление операционной системой и общими ресурсами серверов осуществляется с помощью центральных процессоров Intel Xeon. В сервере NVIDIA DGX H100 используются два 56-ядерных процессора Intel Xeon Platinum 8480C, 3,8 ГГц (Sapphire Rapids). Сервер DGX H200 оснащён двумя центральными процессорами Intel Xeon Platinum 8480CL+ (56 ядер).

Параллельная работа восьми GPU позволяет достичь производительности до 32 петафлопс в вычислениях с пониженной точностью (FP8) и до 8 петафлопс в вычислениях с плавающей точкой (FP16). Объединённая высокоскоростная процессорная память 640 Гбайт и оперативная память 2 Тбайт DDR5 позволяют работать независимо от центров (DC Data Center) для решения небольших задач.

Межсоединение GPU NVIDIA NVLink четвёртого поколения с пропускной способностью 900 Гбайт/с и сетевые интерфейсы NVIDIA ConnectX-7 с поддержкой 400 Гбит/с Ethernet/InfiniBand позволяют поддерживать сверхвысокие скорости обмена данными.

Сервер имеет систему жидкостного охлаждения для GPU и воздушное охлаждение для остальных компонентов. Питается сервер от трёхфазной сети 200–240 В, а также оснащён системами резервного питания (UPS).

Для создания инфраструктуры под крупные модели искусственного интеллекта используются кластеры из сотен систем DGX, объединённых в единую вычислительную среду, такую, например, как DGX SuperPOD. Кластеры представляют собой отдельные помещения, в которых размещены стойки с серверами. В стандартную серверную стойку (42U) можно установить до 5 серверов NVIDIA DGX H100/H200. Кластерные помещения оснащены общей системой питания, охлаждения и другой периферией. Более подробное опи-

сание работы основных перечисленных выше блоков аппаратной части ИИ можно найти на сайте NVIDIA.

Кроме рассмотренных серий NVIDIA H100 и H200 в серверах ИИ широко используются и другие GPU:

- NVIDIA A100 – предшественник H100, всё ещё широко применяется в серверах ИИ с 40/80 Гбайт памяти HBM2e;
- Habana Gaudi – приобретённые Intel чипы, оптимизированные для ИИ рабочей нагрузки;
- AMD Instinct MI250/MI300X – серия GPU от AMD, конкурирующая с NVIDIA. MI300X имеет 192 Гбайт HBM3 памяти;
- Google TPU (v4/v5e) – специализированные тензорные процессоры от Google, используемые в их облачных сервисах;
- NVIDIA L4/L40 – энергоэффективные GPU для инференса с меньшим энергопотреблением;
- Intel Gaudi2/Gaudi3 – ускорители ИИ от Intel, предназначенные для обучения и инференса;
- AWS Trainium/Inferentia – собственные чипы Amazon для обучения и инференса;
- Biren Technology BR100 – высокопроизводительный GPU для ИИ-вычислений с высокой пропускной способностью памяти;
- Moore Threads MTT S3000/S4000 – серия GPU для задач ИИ и научных вычислений;
- Iluvatar CoreX – производительные GPU, используемые для обучения крупных языковых моделей;
- Cambrian/Cambricon MLU – серия ускорителей, специализированных для машинного обучения;
- Huawei Ascend 910/910B – мощные ускорители ИИ от Huawei, разработанные специально для обучения и инференса крупных моделей.

Эти GPU различаются по объёму и типу памяти (от 24 до 192 Гбайт, HBM2e, HBM3); энергопотреблению (от 150 до 700 Вт и выше); производительности (TFLOPS для FP16/BF16/INT8); поддержке различных фреймворков и библиотек ИИ; стоимости.

Часть перечисленных модулей может использоваться для целей инференса крупных моделей. Термин «инференс» (Inference) обозначает процесс использования уже обученной модели искусственного интеллекта для получения результатов или предсказаний на основе новых входных данных. В качестве примера можно привести калькулятор. Сначала калькулятор программируется, то есть его обучают выполнять определённые математические операции. После этого калькулятор можно использовать для инференса, вводя в него определённые числа и совершая с ними математические действия, которым он был обучен.

Во втором столбце схемы аппаратной части (рис. 4) подразумеваются компоненты, которые обеспечивают базовую инфраструктуру, необходимую для работы GPU-кластеров и взаимодействия между ними. Среди наиболее значимых из элементов инфраструктуры можно отметить: сеть высокоскоростных соединений между серверами (100–400 Гбайт); оперативная память серверов (1–2 Тбайт); СХД – системы хранения данных; системы электропитания.

Третий блок на этой схеме описывает инструментарий, обеспечивающий объединение серверов в кластеры и создание необходимой для этого инфраструктуры.

Работа ИИ как программно-аппаратного комплекса представляет собой сложный многоэтапный процесс, требующий значительных вычислительных ресурсов и специализированного программного обеспечения.

Программное обеспечение современных систем искусственного интеллекта представляет собой сложную многоуровневую структуру, напоминающую слоёный пирог. Каждый уровень этой структуры решает свои задачи и обеспечивает работу всей системы в целом. Для удобства понимания разделим эту структуру на три основных уровня: базовый (фундамент), средний (инструментальный) и верхний (пользовательский).

Например, процесс обработки вашего запроса к модели ИИ проходит следующий путь. Верхний уровень принимает ваш запрос через API, проверяет его и направляет в систему

Средний уровень готовит модель к работе, распределяет вычисления между серверами. Базовый уровень выполняет тяжёлые математические вычисления на графических процессорах.

Результат возвращается обратно через все уровни в виде сформулированного ответа.

Базовый уровень обеспечивает самые основные функции, без которых невозможна работа ИИ-системы.

Прежде всего – это операционная система и драйверы. Большинство современных ИИ-систем работает на основе специально оптимизированных

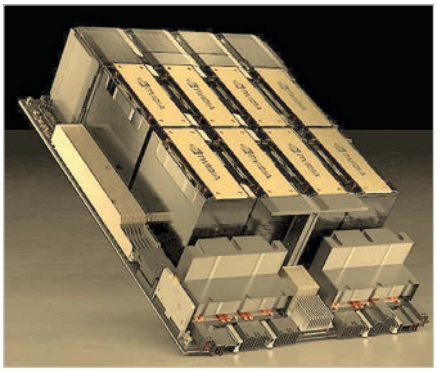

Рис. 5. Сервер NVIDIA DGX H100/H200

версий Linux (Ubuntu Server, CentOS). Эти операционные системы модифицированы для максимальной производительности при работе с графическими процессорами. Компания NVIDIA предлагает собственные оптимизированные версии операционных систем.

Платформы для параллельных вычислений, такие как, например, CUDA от NVIDIA, представляют собой специальный программный интерфейс, позволяющий использовать графические процессоры для неграфических вычислений. Существуют также аналоги от других производителей: ROCm (для AMD), OneAPI (Intel), CANN (Huawei).

Средний уровень содержит инструменты, с которыми работают разработчики ИИ-систем, такие как фреймворки глубокого обучения. Это программные платформы, упрощающие создание и обучение нейронных сетей. Среди наиболее популярных можно выделить: PyTorch, который отличается гибкостью и удобством экспериментирования; TensorFlow, предлагающий как простой интерфейс через Keras, так и возможности тонкой настройки; JAX, специализирующийся на научных вычислениях; MindSpore, оптимизированный для оборудования Huawei.

Для эффективной работы с крупными моделями, содержащими миллиарды параметров, созданы специальные библиотеки: DeepSpeed (Microsoft), оптимизирующая использование памяти; Megatron-LM, которая предназначена для распределённых вычислений; vLLM, ускоряющая инференс языковых моделей.

Системы оркестрации координируют работу множества серверов, объединённых в единую систему: Kubernetes, использующийся как стандартное решение для управления контейнерами; Slurm, традици-

онно популярный в научных кластерах; Ray, предназначенный в качестве универсального фреймворка для распределённых приложений.

Верхний уровень обеспечивает доступ к возможностям ИИ для конечных пользователей и внешних систем.

Серверные API верхнего уровня представляют собой интерфейсы, через которые пользователи и приложения взаимодействуют с моделями ИИ. Из наиболее популярных можно выделить FastAPI и Flask. Это лёгкие и быстрые решения для создания API. Утилита gRPC обеспечивает более эффективную передачу данных при высоких нагрузках.

Верхний уровень также содержит программы для поддержки инференса, такие как Triton Inference Server и TorchServe.

Системы кэширования и балансировки нагрузки Redis, Memcached и NGINX обеспечивают быстрый доступ при больших загрузках к часто используемым серверам.

Кроме того, верхнему уровню принадлежат также утилиты мониторинга и логирования, которые отслеживают работу системы и помогают выявлять проблемы, например, Prometheus, Grafana, Elasticsearch, Logstash, Kibana и другие.

Перечисленные программные продукты были приведены здесь только в качестве известных примеров и далеко не полностью представляют всю линейку программных продуктов, задействованных в современных ИИ. Более подробное описание принципов работы современных мощных моделей ИИ можно посмотреть в детальном обзоре [15].

Цель этого раздела заключалась в том, чтобы показать, что такая многоуровневая архитектура позволяет эффективно управлять всем жизненным циклом больших языковых моделей, обеспечивая оптимальное использование вычислительных ресурсов и высокое качество обслуживания конечных пользователей. При этом каждый уровень развивается в значительной степени независимо, что позволяет внедрять инновации без перестройки всей системы.

Ведущие компании – разработчики искусственного интеллекта

Из огромного количества появившихся в настоящее время компаний, занимающихся разработками ИИ, трудно выделить те, которые занимают первые места в этом направлении. Тем не менее, основываясь на мнении наиболее авторитетных экспертов в области искусственного интеллекта, можно говорить о тех компаниях, чья продукция представляется наиболее перспективной.

На основе прогнозов развития ИИ, сделанных такими гигантами, как Google [16]; IBM IBV [17, 18]; Gartner [19]; Statista [20]; AlphaSens [21]; CB Insights [22], мы постарались выделить те компании, которые действительно занимают лидирующие направления в разработках наиболее важных приложений ИИ.

Таблица 1. Общепринятая в англоязычной литературе аббревиатура базовых направлений искусственного интеллекта

ABI	Agent-Based AI	Агентные системы ИИ
AGM	Audio Generative Model	Генерация аудио
AGI	Artificial General Intelligence	Сильный искусственный интеллект
AIS	AI Supercomputer	ИИ на базе суперкомпьютера
ATS	Autonomous Transport Systems	Автономные транспортные системы
AR	Accuracy Rewards	Награды за точность решения
BCI	Brain-Computer Interfaces	Интерфейсы мозг–компьютер
BDS	AI Bias Detection Systems	Системы обнаружения предвзятости ИИ
CSD	Cold-Start Data	Начальный анализ на основе ранее накопленного опыта
CV	Computer Vision	Компьютерное зрение
CV&IG	Computer Vision & Image Generation	Компьютерное зрение и генерация изображений
DGX	Data Center GPU eXtreme	Дата-центр Nvidia
GAN	Generative Adversarial Networks	Генеративные состязательные сети
GPU	Graphics Processor Unit	Графические процессоры GPU
GRPO	Group Relative Policy Optimization	Оптимизация групповой относительной политики
IGM	Image Generation Models	Модели генерации изображений
IFT	AI Fairness Tools	Инструменты обеспечения справедливости ИИ
FR	Format Rewards	Награды за чётко структурированный вывод
LLM	Large Language Models	Большие языковые модели
MFR	Multifunctional Robots	Многофункциональные роботы
NGA	New Generation AI	Новое поколение ИИ
POD	Performance Optimized Datacenter	Модульные кластеры Nvidia
PPO	Proximal Policy Optimization	Оптимизация ближайшей политики
PRL	Pure Reinforcement Learning	Чистое обучение с подкреплением
RIP	Responsible AI Practices	Ответственные практики ИИ
RLHF	Reinforcement Learning from Human Feedback	Обучение с обратной связью от человека
RBR	Rule-Based Rewards	Награды на основе правил RBR
S2T	Speech-to-Text	Модели преобразования голоса в текст
SAI	Specialized AI systems	Специальные системы ИИ
SLM	Small Language Models	Малые языковые модели
SFT	Supervised Fine-Tuning	Тонкая настройка с помощью человека-супервизора
T2I	Text-to-Image	Модели преобразования текста в изображение
T2S	Text-to-Speech	Модели преобразования текста в голос
TPU	Tensor Processing Units	Тензорный процессор
VGM	Video Generative Model	Генерация видео

В этой статье не рассматриваются российские большие модели ИИ, которым посвящено достаточно большое количество статей в Интернете. Подробный обзор российских компаний, которые занимаются созданием решений на основе генеративного ИИ, а также сопутствующих инструментов, доступен на сайте [23].

Поскольку в этой статье цитируются в основном источники ведущих мировых компаний, ниже использована общепринятая в англоязычной литературе аббревиатура, расшифровка которой приведена в табл. 1.

Ниже коротко рассмотрены только несколько наиболее крупных и вли-

ятельных мировых компаний, разрабатывающих и производящих модели искусственного интеллекта (AI Companies – AIC), среди которых, прежде всего, следует отметить: OpenAI, Google AI, Meta AI, Microsoft AI, Anthropic, Core AI (подразделение Microsoft), Stability AI, xAI, Amazon AI Web Services, Midjourney.

По всей видимости, одной из первых специализированных компаний, которые стали интенсивно инвестировать в направление ИИ, можно считать Facebook Artificial Intelligence Research – FAIR, созданную в концерне Facebook в 2013 году. Следует отметить, что первым генеральным директором FAIR был Ян Лекюн, лауреат премии Тьюринга, профессор Нью-Йоркского университета.

Компания Meta AI появилась в 2021 году после того, как компания FAIR была переименована в результате ребрендинга Facebook в Meta Platforms Inc. Научно-исследовательское подразделение Meta AI фокусируется на разработке открытых моделей ИИ, машинного перевода, компьютерного зрения и других инновационных технологий [24].

В 2015 году была основана компания OpenAI. Сопредседателями организации стали Илон Маск и Сэм Альтман. Первоначально OpenAI была создана как некоммерческая организация с целью разработки и продвижения искусственного интеллекта, который будет полезен для всего человечества. Однако в 2019 году компания перешла на коммерческую бизнес-модель «ограниченной прибыли» [25].

Компания DeepMind Technologies – британская компания, специализировавшаяся на разработке искусственного интеллекта. Она была основана в 2010 году в Лондоне. В 2014 году компанию приобрёл концерн Google, и она была переименована в DeepMind.

В 2023 году DeepMind объединилась с подразделением Google Brain и стала частью Google DeepMind. Основное направление – разработка моделей искусственного интеллекта, включая разработку систем, способных решать математические задачи и создавать новые материалы с уникальными свойствами.

Другое структурное отделение корпорации Google, созданное в 2017 году, получило название Google AI. Эта самостоятельная компания разрабатывает различные проекты и технологии, такие как, например, TensorFlow, LaMDA и Sycamore. Подразделение Google AI занимается также другими проектами: от машинного обучения и компьютерного зрения до медицинских приложений и этических аспектов ИИ [26].

Кроме того, во всём мире известна облачная платформа со свободным доступом Google-AI-Studio [27].

Компания Anthropic была основана в 2021 году бывшими сотрудниками OpenAI, среди которых был Дарио Амодей, ранее занимавший должность вице-президента по исследованиям в OpenAI. Его сестра Даниэла Амодей стала президентом Anthropic. Сегодня Anthropic фокусируется на разработке безопасных и интерпретируемых систем ИИ. Большое внимание компания уделяет этическим и экологическим проблемам, связанным с ИИ. Она имеет статус Public Benefit Corporation – PBC [28].

Корпорация Microsoft начала активно заниматься проблемами ИИ с разработок в области машинного обучения и когнитивных сервисных приложений только в конце 2000-х.

В марте 2023 года было официально анонсировано создание отдельного подразделения Microsoft AI, целью которого стали вопросы развития рынка продаж потребительских товаров на основе искусственного интеллекта [29].

В начале 2025 года корпорация Microsoft объявила о создании нового подразделения Core AI Platform and Tools.

В состав нового отделения включены такие группы разработчиков, как, например:

- Development Division (разработка инструментов для программистов);
- AI Platform (создание базовых технологий и инфраструктуры для работы с искусственным интеллектом);
- AI Supercomputer (разработка высокопроизводительных вычислительных систем для обучения и запуска крупных ИИ-моделей);
- AI Agentic Runtimes (разработка автономных ИИ-систем, способных выполнять задачи самостоятельно);
- Engineering Thrive (оптимизация инженерных процессов и повышение их эффективности).

Объединение этих команд позволяет создать более интегрированный подход к разработке ИИ-платформ и инструментов, которые могут быть использованы как внутри компании, так и предложены внешним разработчикам и клиентам.

Кроме того, Microsoft продолжает тесное сотрудничество с OpenAI и активно инвестирует в развитие ИИ-технологий [30].

Компания Stability получила широкую известность в последние годы и вошла в список ведущих вендоров ИИ благодаря своим генеративным моделям для создания изображений и аудио. Так, в перечне продукции Stability есть ИИ различных классов, таких как: большая языковая модель; генерация изображений; модели для работы с аудио и генерации кода [31].

Ещё одна относительно новая ИИ, основанная на базе xAI по инициативе Илона Маска, появилась в 2023 году. По размаху и вложениям эта компания, вероятно, занимает первое место. Она разработала одну из самых мощных сейчас LLM моделей Grok, которая имеет прямой доступ к облачному сервису X, ранее известному как Twitter.

Необходимо сказать несколько слов о том, что представляет собой аппаратная часть Grok.

Например, в моделях ИИ Grok используется линейка высокопроизводительных графических процессоров H100 от NVIDIA, разработанных специально для задач искусственного интеллекта и глубокого обучения. Дата-центры Data Center GPU eXtreme – DGX объединены в модульные кластеры Performance Optimized Datacenter – POD, предназначенные для масштабируемых вычислений.

Например, для обучения модели Grok-1.5 используется кластерная система, содержащая до 5000 графических процессоров NVIDIA H100. Каждый H100 имеет 80 Гбайт памяти HBM и обеспечивает до 3,958 петафлопс вычислительной мощности в FP8 [33].

По некоторым данным, xAI использует инфраструктуру Oracle Cloud, а также, возможно, собственные вычислительные мощности [34].

Для размещения 5000 GPU необходима общая площадь порядка 150–200 м² серверного пространства (с учётом проходов, систем охлаждения и электроснабжения) [35].

Учитывая, что средняя стоимость одного GPU NVIDIA H100 находится в диапазоне от 25 до 30 тысяч долларов США, стоимость только GPU составит $125–150 млн. Кроме того, серверное оборудование, сетевая инфраструктура дополнительно составят около $40–60 млн. Системы охлаждения, электропитания и резервного копирования – это ещё $20–30 млн. В результате суммарная стоимость оборудова-

Рис. 6. Кластерная система GPU серверов в ИИ Grok [37]

ния может превысить $240 млн. Также нужно учесть дополнительные расходы, связанные с потреблением электроэнергии. Если учесть, что в сумме такое оборудование потребляет около 20 МВт, годовые затраты на электроэнергию составят ещё $15–20 млн. Нужно не забыть расходы на обслуживание инфраструктуры – $10–15 млн в год [36].

По оценкам экспертов, для обучения последней модели Grok 3 было задействовано около 100 000 графических процессоров (GPU NVIDIA H100). Стандартная серверная система DGX NVIDIA H100 включает в себя 8 GPU H100. Поэтому для размещения 100 000 GPU H100 потребуется 12 500 систем DGX H100 (рис. 6) [37].

Своеобразным сюрпризом в 2024 году стала китайская компания DeepSeek, которая является одним из ключевых китайских игроков на рынке ИИ. Она была основана в 2023 году бывшими сотрудниками из Baidu Янь Цзюнем (бывший технический вице-президент Baidu) и Цао Яндунем (экстехнический директор Baidu) [38]. Эта компания выделяется среди китайских компаний, занимающихся разработкой моделей ИИ. Неслучайно в названии компании обыгрываются слова «Deep Seek» – «глубинный поиск» (рис. 7).

Используя нетрадиционный подход как в аппаратной, так и в программной частях своей разработки, DeepSeek сумела создать модель, которая по производительности не уступает, например, LLM OpenAI, однако стоит в несколько раз дешевле [39].

Добавим, что DeepSeek – один из немногих китайских разработчиков, выпускающих как проприетарные, так и открытые модели ИИ [40].

Модель DeepSeek Coder 33B была специально разработана для лучшего понимания и генерации программного кода [41].

Кардинально отличается аппаратная часть DeepSeek от американских и других западных производителей. Из крайне скупых описаний состава аппаратной части можно предполагать, что DeepSeek, по оценкам экспертов, использует, скорее всего, последнюю модель китайского специализированного процессора Huawei Ascend 910C. Однако конкретных упоминаний этого факта нет, по крайней мере, в англоязычной литературе. Можно лишь сказать, что этот процессор, разработанный Huawei специально для задач ИИ, основан на архитектуре Da Vinci, оптимизированной для обработки больших объёмов данных и выполнения сложных вычислений, необходимых для работы с глубокими нейронными сетями. Наличие встроенной памяти снижает необходимость частого обращения к внешним источникам данных и ускоряет обработку больших объёмов данных. Для обучения моделей уровня DeepSeek R1, по-видимому, используется кластер из сотен или тысяч GPU/TPU. Например, тренировка на Nvidia H800 могла проходить на кластере из примерно тысячи единиц, а вывод – на аналогичном количестве Ascend 910C [42, 43].

Ориентировочная цена одного чипа Ascend 910C может составлять примерно $5000, что в несколько раз дешевле Nvidia H100. Поэтому несмотря на то, что производительность китайского процессора Ascend 910C достигает всего 60% от производительности американского H100, он значительно выигрывает в цене [44].

Большинство крупных компаний развивает различные направления ИИ одновременно.

Вместе с тем существует определённая специализация компаний – разработчиков ИИ по отдельным направлениям:

- OpenAI: продвинутые языковые и мультимодальные модели;
- Anthropic: безопасность и этичность ИИ;
- Meta AI: открытые модели и исследования;
- Google: широкий спектр от фундаментальных исследований до практических приложений;
- Microsoft: интеграция ИИ в продукты и разработка инструментов;
- xAI (Grok): акцент на максимальной информационной открытости, нестандартном мышлении и встроенной актуальной информации из Интернета;
- DeepSeek: специализированные модели для программирования и математики с фокусом на китайский и мировой рынок, баланс между открытыми и проприетарными решениями.

Языковые модели искусственного интеллекта

С помощью обучающих программ большие генеративные модели, кроме того, что способны решать сложнейшие технические вопросы со скоростью обработки миллионов операций в секунду, могут также писать музыку, стихи, генерировать изображения, поддерживать диалог практически на всех языках мира (рис. 8).

Языковые модели искусственного интеллекта (Language Models – LM), являющиеся частью общего класса Generative Language Models – GLM, предназначены для обработки естественного языка. Эти модели могут распознавать, переводить, предсказывать или генерировать текст или другой контент, включая видео и изображения.

Различают два типа языковых моделей – большие (Large Language Models – LLM) и малые (Small Language Models).

Современные LLM обычно имеют от 70 млрд параметров и выше. Например: GPT-4 – примерно от 1 до 1,8 трлн; Claude 3 – более 1 трлн; Gemini Ultra – более 100 млрд; GPT 3,5–175 млрд.

Классификация моделей LLM основывается не только на количестве параметров, но также учитывает такие функциональные возможности, как требования к вычислительным ресурсам; возможность локального запуска; эффективность использования ресурсов; специализация конкретных задач.

В качестве примера наиболее известных больших языковых моделей LLM можно привести следующие:

- GPT-4 от OpenAI (ChatGPT) [46];
- ChatGPT от OpenAI [47];
- Gemini от Google [48];
- Claude от Anthropic [49];
- LLaMA 2 от Meta [50];
- Grok от xAI [51];
- PaLM/PaLM 2/CodeGemma от Google [52];

Рис. 7. DeepSeek – глубинный поиск [39]

- Stability AI [53];
- и другие.

Из новых моделей можно выделить Claude 3.7, Grok 3 и Gemini 2.0.

Последняя версия Claude 3.7 Sonnet была выпущена в феврале 2025 года. Это часть семейства моделей Claude 3, которое включает также Claude Haiku и Claude Opus. Увеличенное окно контекста (двести тысяч токенов) демонстрирует улучшенное понимание контекста, более связные и релевантные ответы, а также учёт нюансов и подтекста вопросов. Кроме того, модель Claude 3.7 способна лучше справляться со сложными, многоступенчатыми задачами, требующими более глубокого анализа. Например, Claude 3.7 может без ошибок генерировать инструкции для настройки таких сложных электронных систем, как серверы с ИИ-ускорителями [54].

Одна из последних LLM моделей Google DeepMind Gemini 2.0 появилась в феврале 2025 года. Кроме того, что для продвинутых пользователей появилась возможность решения сложных задач через API Gemini в Google AI Studio, обновлённый вариант ИИ с Flash Gemini 2.0 стал доступным на рабочем столе и мобильных устройствах для всех пользователей приложения Gemini.

Внедрение LLM Gemini 2.0 в управление IoT «умного дома» позволит существенно упростить процессы использования ИИ в таких новых развивающихся направлениях, как AI-In-Everything – AIIE.

Например, ИИ сможет быть полезным в тех случаях, когда нужно будет генерировать код, компилировать и адаптировать его для ПО умного термостата, который должен отвечать на сложные голосовые запросы типа «почему так жарко, сделай похолоднее» и т.д. [55].

Мультимодальная модель ИИ OpenAI GPT-4o (Omni), выпущенная в мае 2024 года, обладает улучшенными возможностями обработки текста, изображений, аудио и видео в реальном времени, позволяя вести естественные разговоры со способностью переключаться между модальностями. Модель GPT-4o значительно превосходит предыдущие разработки по скорости и точности, особенно в задачах распознавания и анализа речи [56].

Большая языковая модель Grok 3, появившаяся в феврале 2025 года, предназначена для обработки гигантских объёмов различных типов данных в виде текста и изображений [57].

Из основных отличительных свойств Grok 3 можно выделить такие, например, как: доступ к сетям Интернет в реальном масштабе времени; режим Think; функция DeepSearch. Режим Think позволяет разбивать сложные проблемы на простые и решать их последовательно шаг за шагом, уточняя ответы на каждой следующей итерации. В отличие от статических моделей, Grok имеет функцию DeepSearch, которая даёт возможность извлекать информацию в реальном времени из Интернета и BDB (Big Database) и генерировать ответы, отражающие последние достижения науки и техники. Обучение Grok 3, реализованное на суперкомпьютере xAI Colossus с более чем ста тысячами графических процессоров Nvidia (точные оценки неизвестны), а также контекстное окно, превышающее, по гипотетическим оценкам, миллион токенов, обеспечивают непрерывную поддержку чатов в течение длительного времени.

Важно подчеркнуть, что Grok 3 – это не просто генерация текста, а глубокий анализ проблемы. Так, например, этот ИИ может быть успешно использован при разработке чипов в плане создания кодов оптимизации, а также для общения с периферийными интеллектуальными устройствами [58].

Семейство крупных языковых моделей DeepSeek LLM представлено разными продуктами с размерами от 7 млрд до 67 млрд параметров. Модель DeepSeek 67B демонстрирует производительность, сравнимую с GPT-3.5. Она обучена на массивном датасете из более чем 2 трлн токенов.

Модель DeepSeek Coder 33B – это SLM, дополнительно обученная на больших корпусах кода и технической документации. Базовая архитектура взята от DeepSeek LLM, но модель прошла специализацию для лучшего понимания и генерации программного кода. Она была дополнительно обучена на триллионах токенов кода из различных репозиториев.

Модель DeepSeek Math 67B также является LLM, специализированной на математических задачах. Она построена на основе базовой LLM-архитектуры DeepSeek, но прошла дополнительное обучение на математических датасетах, включая задачи, формулы, доказательства и вычисления.

Коренное отличие DeepSeek от других моделей заключается в их новом подходе к обучению.

Рис. 8. Современные большие модели ИИ могут писать музыку, стихи, генерировать изображения [45]

Практически все большие современные модели, такие как OpenAI ChatGPT, используют для обучения метод Reinforcement Learning from Human Feedback – RLHF. В этом случае модель сначала обучается на больших объёмах данных (предобучение), а затем проходит этап тонкой настройки с использованием человеческой обратной связи. Человек оценивает ответы модели как «хорошо» или «плохо», и на основе этих оценок модель оптимизируется через алгоритмы усиленного обучения, такие как Proximal Policy Optimization – PPO. Это помогает модели лучше соответствовать человеческим ожиданиям и предпочтениям. В отличие от классического обучения на данных, здесь ИИ «экспериментирует», а человек говорит, что хорошо, а что плохо, точно так, как это делает тренер для спортсмена.

Основные этапы обучения LLM моделей методом RLHF приведены ниже (рис. 9).

1. Инициализация. Определяется задача, которую должен освоить ИИ-агент, и оценивается соответствующая функция вознаграждения.

2. Сбор и предварительная обработка данных. Отбираются решения задач экспертами. Эти демонстрации служат примерами, на основе которых обучается ИИ-агент. Полученные данные обрабатываются и преобразуются в формат, подходящий для обучения.

3. Начальное обучение политики поведения. Агент учится имитировать поведение экспертов на основе собранных данных.

4. Применение политики. Взаимодействие ИИ-агента с окружающей средой с использованием выученной политики.

5. Обратная связь. Эксперты предоставляют обратную связь о действи-

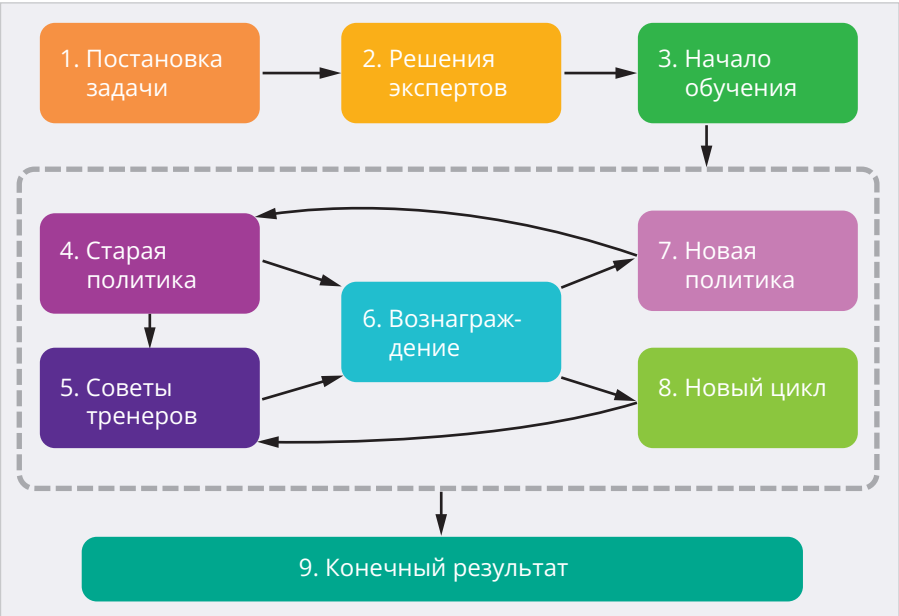

Рис. 9. Упрощённая схема обучения модели ИИ с подкреплением на основе обратной связи с человеком – RLHF

ях агента. Эта обратная связь может быть бинарной оценкой (хорошо или плохо) или более детализированной.

6. Обучение модели получения вознаграждения. Использование обратной связи от людей для создания модели получения вознаграждения, которая отражает предпочтения экспертов.

7. Обновление политики поведения на основе усвоенной агентом модели вознаграждения.

8. Итеративный процесс: этапы 4–7 повторяются итеративно, при этом ИИ-агент каждый раз совершенствует свою политику на основе демонстраций и обратной связи.

9. Процесс RLHF продолжается до тех пор, пока производительность агента не достигнет удовлетворительного уровня или пока не будет достигнут заранее определённый критерий остановки.

Следует отметить, что в варианте RLHF требуются огромные объёмы данных обратной связи, основанные на опыте миллионов людей. В противоположность этому подход DeepSeek (R1-Zero) полностью исключает человеческую обратную связь.

В альтернативном варианте методика DeepSeek (R1-Zero) предлагает коренное отличие в методе обучения ИИ. Компания используют собственный алгоритм Group Relative Policy Optimization – GRPO, который не требует отдельной модели-критика (как в PPO) и оптимизирует модель на основе сравнения группы сгенерированных ответов. При этом исключаются этап Supervised Fine-Tuning – SFT и человеческая обратная связь на этапе постобучения, присущие RLHF. Вместо того чтобы использовать оценку человека, DeepSeek-R1-Zero обучается с помощью метода Pure Reinforcement Learning – RL, основанному на правильно-ориентированных наградах Rule-Based Rewards. Так, например, модель получает награду, которая называется Accuracy Rewards, если она абсолютно правильно решает математическую задачу. Другую награду Format Rewards модель получает, например, за чётко структурированный вывод. Каждая награда имеет свой вес. Таким образом, регламентируется каждое действие, и модель DeepSeek-R1-Zero «самообучается» без прямого вмешательства человека, что коренным образом отличается от метода RLHF.

Стоит отметить, что полная линейка DeepSeek-R1 не ограничивается только «чистым RL». Если DeepSeek-R1-Zero представляла собой эксперимент с полным исключением SFT и человеческой обратной связи, то финальные последующие модели включали комбинированные процессы.

Так, модель DeepSeek-R1 включает многоступенчатый процесс обучения:

- этап «холодного старта» с использованием данных (Cold-Start Data) для начальной настройки;
- основной этап RL (как в R1-Zero);
- дополнительная тонкая настройка с использованием данных, сгенерированных моделью, и повторный RL для улучшения читаемости и согласованности.

Подход DeepSeek, который учит ИИ думать самостоятельно, без подсказок от людей, является более автономным и менее трудозатратным, в отличие от трудоёмкого сбора информации, практикуемого OpenAI [59].

Понятие «большие языковые модели» неразрывно связано с так называемыми нейронными сетями – НС (Neural Networks – NN), представляющими собой вычислительные системы, основанные на многослойной базе сети узлов, созданных по аналогии с мозгом человека [60] (рис. 10).

Современные языковые модели анализируют миллиарды текстов: от шекспировских пьес до научных статей и программного кода, обнаруживая скрытые языковые закономерности. Это позволяет им понимать намерения человека даже при неточных формулировках, генерировать практически неотличимый от созданного человеком творческий контент, а также решать задачи, требующие сложного логического мышления.

В реальной жизни большие модели (LLM) уже сегодня широко используются для автоматизации работы с рутинной документацией, мгновенных переводов на большинство языков мира с сохранением контекстных нюансов. Они могут играть роль виртуальных секретарей, способных поддержать содержательную дискуссию практически на любую тему.

Кроме LLM существуют и другие модели, оперирующие с небольшим числом параметров (меньше 15 млрд) и оптимизированные для выполнения специфических задач с меньшими вычислительными затратами. Они получили название «малые языковые модели» (Small Language Models – SLM). Эти модели представляют собой упрощённые версии больших языковых моделей.

Мультимодальные и перцептивные модели представляют класс ИИ, который характеризуется способностью инте-

Рис. 10. Понятие «большие языковые модели» неразрывно связано с так называемыми нейронными сетями [61]

грировать разные типы данных (текст, изображения, аудио, видео) в единый целостно воспринимаемый контент.

В отличие от пассивных моделей ИИ, развиваются так называемые агентные системы ИИ (Agent-Based AI – ABAI), которые способны действовать самостоятельно, проявлять инициативу, учиться на своём опыте и адаптироваться к меняющимся условиям.

Отдельный класс составляют модели, предназначенные для инструментальной поддержки процесса разработки программного обеспечения ИИ.

Эти и другие вопросы развития ИИ будут рассмотрены в следующих частях.

Список литературы

1. Немного об искусственном интеллекте. URL: https://lawinrussia.ru/nemnogo-ob-iskusstvennom-intellekte/.
2. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/Main_Page.
3. Национальная стратегия развития ИИ на период до 2030 года. URL: https://bit.ly/4hVxJq4.
4. Конференция «Путешествие в мир искусственного интеллекта». URL: http://kremlin.ru/events/president/news/75830.
5. «GigaChat», СБЕРБАНК. URL: https://giga.chat/.
6. Kandinsky 3.1, Sberbank. URL: https://www.sberbank.com/promo/kandinsky/.
7. Yandex GPT-2, Yandex. URL: https://ya.ru/ai/gpt-2.
8. «Телематика», ComNews. URL: https://www.comnews.ru/content/235339/2024-09-23/2024-w39/1010/bespilotnye-gruzoviki-poekhali-pomoschyu-cifrovogo-dvoynika-trassy-m-11-koncerna-telematika.
9. AI.T-bank, Центр искусственного интеллекта Т-Банка. URL: https://ai.tbank.ru/.
10. РБК Тренды. URL: https://bit.ly/4i292s5.
11. Искусственный интеллект в России – 2023: тренды и перспективы. URL: https://yakov.partners/upload/iblock/c5e/c8t1wrkdne5y9a4nqlicderalwny7xh4/20231218_AI_future.pdf.
12. ООО «Зерокодер». URL: https://zerocoder.ru/.
13. Перечень поручений по итогам конференции «Путешествие в мир искусственного интеллекта». URL: http://kremlin.ru/acts/assignments/orders/76076/print.
14. Искусственный интеллект. URL: https://bit.ly/4h4sx1S.
15. The rise of Large Language Models: from fundamentals to application. URL: https://www.managementsolutions.com/sites/default/files/minisite/static/72b0015f-39c9-4a52-ba63-872c115bfbd0/llm/pdf/rise-of-llm.pdf.
16. AI Business Trends 2025, report. Google. URL: https://cloud.google.com/resources/ai-trends-report.
17. 5 Trends for 2025. IBM IBValue. URL: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/business-trends-2025.
18. 2025 Global Outlook for Banking and Financial Markets. URL: https://www.ibm.com/thought-leadership/institute-business-value/.
19. Use the 2025 strategic technology trends to shape the future with responsible innovation. Gartner. URL: https://www.gartner.com/en/articles/top-technology-trends-2025.
20. AI-trends-2025. Statista. URL: https://www.statista.com/free-content/thank-you/ai-trends-2025-cs.
21. Generative AI for Million-Dollar Decisions. AlphaSense. URL: https://bit.ly/43dSby3.
22. AI super analyst for market intelligence. CBINSIGHTS. URL: https://www.cbinsights.com/.
23. Карта рынка российского GenAI и сопутствующих продуктов. Хабр. URL: https://habr.com/ru/articles/879622/.
24. Meta AI. URL: https://ai.meta.com/.
25. OpenAI. URL: https://openai.com/.
26. Google AI. URL: https://ai.google/.
27. Google-AI-Studio. URL: https://blog.click.ru/glossary/google-ai-studio/.
28. Anthropic. URL: https://www.anthropic.com/.
29. Microsoft AI. URL: https://www.microsoft.com/en-us/ai.
30. Introducing Core AI Platform and Tools. URL: https://blogs.microsoft.com/blog/2025/01/13/introducing-core-ai-platform-and-tools/.
31. Stability AI. URL: https://stability.ai/.
32. The NVSWITCH fabric that is the hub of the DGX H100 SUPERPOD. URL: https://www.nextplatform.com/2022/03/23/nvidia-will-be-a-prime-contractor-for-big-ai-supercomputers/.
33. NVIDIA. URL: https://www.nvidia.com/en-us/on-demand/session/gtcspring23-se52203/.
34. NVIDIA H100 Tensor Core GPU Architecture. URL: https://developer.nvidia.com/blog/nvidia-h100-tensor-core-gpu-architecture/.
35. NVIDIA DGX H100 Technical Specifications. URL: https://www.nvidia.com/en-us/data-center/dgx-h100/.
36. Gartner о рынке ИИ-инфраструктуры. URL: https://www.gartner.com/en/documents/4116090.
37. Colossus Inside the 100K GPU xAI Cluster that Supermicro Helped Build for Elon Musk. URL: https://www.servethehome.com/inside-100000-nvidia-gpu-xai-colossus-cluster-supermicro-helped-build-for-elon-musk/.
38. DeepSeek. URL: https://www.deepseek.com/.
39. DeepSeek – глубинный поиск. URL: https://bit.ly/4i1UZ5J.
40. GitHub. DeepSeek AI. URL: https://github.com/deepseek-ai.
41. DeepSeek Coder 33B. URL: https://huggingface.co/deepseek-ai/deepseek-coder-33b-instruct.
42. Huawei Ascend GPU. URL: https://e.huawei.com/en/products/computing/ascend.
43. Huawei Cloud. URL: https://www.huaweicloud.com/intl/en-us/product/flexus.html.
44. Huawei Enterprise. URL: https://e.huawei.com/eu/.
45. Imagine, Forbes. URL: https://bit.ly/3DbjcYn.
46. GPT-4 от OpenAI. URL: https://openai.com.
47. ChatGPT от OpenAI. URL: https://chatgpt.com/.
48. Gemini от Google. URL: https://deepmind.google/technologies/gemini/.
49. Claude от Anthropic. URL: https://www.anthropic.com.
50. LLaMA 2 от Meta. URL: https://ai.meta.com/llama/.
51. Grok от xAI. URL: https://x.ai/.
52. PaLM/PaLM 2/CodeGemma от Google. URL: https://ai.google/get-started/our-models/.
53. Stability AI Models. URL: https://stability.ai/.
54. Claude 3.7 Sonnet. URL: https://www.anthropic.com.
55. Gemini 2.0 API. URL: https://aistudio.google.com/prompts/new_chat?model=gemini-2.0-pro-exp-02-05&pli=1.
56. GPT-4o (Omni). URL: https://openai.com/index/hello-gpt-4o/.
57. Grok 3. URL: https://grok.com/.
58. Grok 3 Signup. URL: https://accounts.x.ai/sign-up?redirect=grok-com&theme=light.
59. What is reinforcement learning from human feedback (RLHF). URL: https://bdtechtalks.com/2023/01/16/what-is-rlhf/.
60. What are large language models (LLMs)? URL: https://www.elastic.co/what-is/large-language-models.
61. Искусственный интеллект: генерация изображения. URL: https://goo.su/MKpDW.