

Возможности использования генеративного ИИ для обеспечения технологического суверенитета российского ТЭК

Генеративный ИИ обладает огромным потенциалом, в его развитие в мире вкладываются десятки миллиардов долларов. Сложность и качество больших языковых моделей (БЯМ) растут, хотя они пока плохо справляются с задачами, требующими многоступенчатых рассуждений и математических действий.

В перспективе БЯМ научатся решать сложные многоуровневые инженерные задачи и создавать новые технологии для российского нефтегаза, превратившись в прикладной инструмент для обеспечения его технологического суверенитета. Суммарный отраслевой эффект от использования генеративного ИИ может составить 343 млрд руб. в год за счет повышения производительности труда и эффективности производства.

Лучшие зарубежные БЯМ сегодня не отвечают на конкретные отраслевые вопросы, во многом из-за отсутствия в их корпусе данных с отраслевой и страновой спецификой. Технология RAG совместно с функционалом «умного поиска» улучшает качество ответов за счет использования актуальных данных.

Для создания отраслевой БЯМ необходимо объединение компетенций отрасли и ИТ-компаний на единой платформе, разделение финансирования и обмен информацией. При этом развитие этой платформы должно происходить поэтапно, с постепенным увеличением сложности решаемых задач.

Предисловие



ВЫГОН Григорий Вадимович
Управляющий директор, к.э.н.

Сегодня мы наблюдаем рост конкуренции на глобальных рынках нефти и газа. Происходит это из-за быстрого развития технологий, в том числе цифровых, во всех сегментах производства и потребления энергоресурсов. Для российского ТЭК, живущего в условиях санкционных ограничений, переориентации экспортных потоков и энергоперехода, вопрос обеспечения технологического суверенитета становится экзистенциальным.

Доля импортного высокотехнологичного оборудования и материалов, используемых в разведке, добыче, переработке углеводородного сырья, транспортировке нефти, газа и производимых из них продуктов все еще велика. Из списка критического оборудования, от которого зависит отрасль, достаточно назвать турбины, компрессоры и теплообменники для производства СПГ, газовозы для его транспортировки, установки гидрокрекинга и каткрекинга в переработке, пиролиза в нефтехимии и т.п. Ежегодные потребности отечественного нефтегаза в импортозамещении для текущих и заявленных проектов оцениваются нами в 200-300 млрд руб. до 2035 г.

Одним из эффективных способов реагирования на эти вызовы может выступить генеративный ИИ. Большие языковые модели умеют классифицировать и резюмировать информацию, строить простые логические выводы. В ближайшем будущем им по силам станут сложные многоуровневые инженерные задачи – проектирование процессов, оборудования и инфраструктуры, а также создание новых продуктов и технологий. В результате у российских ученых и инженеров появится эффективный прикладной инструмент для интенсификации импортозамещения.

Для создания отраслевой платформы на базе БЯМ необходимо создать консорциум заинтересованных нефтегазовых компаний, готовых инвестировать в развитие модели для решения больших задач. Комплексный проект по развитию генеративного ИИ, в котором участвуют профильные ФОИВ, нефтегазовые компании и их подрядчики, IT-разработчики и отраслевые эксперты поможет российскому нефтегазу значительно продвинуться в обеспечении технологического суверенитета и конкурентоспособности на мировых рынках энергоносителей.

Ключевые выводы (1/4): GenAI обладает огромным потенциалом, в его развитие вкладываются 10-ки млрд долл., сложность и качество БЯМ растут

Обзор глобальных трендов развития больших языковых моделей

1. **Генеративный ИИ** является нейросетью с глубоким машинным обучением на неструктурированной информации. В отличие от классического ИИ, GenAI обучается на гораздо большем объеме (на ~3 порядка) мультимодальных данных и генерирует новый контент по запросам пользователя.
2. **Большие языковые модели (БЯМ)** являются частным случаем нейронных сетей, специализирующиеся на задачах обработки естественного языка и обладающие способностью генерировать текст и изображения на основе входных данных.
3. Каждый месяц появляются **8-10 новых БЯМ**, при этом количество мега-моделей (~2 трлн параметров) растет, лидируют американские команды **OpenAI, Google, Anthropic**.
4. БЯМ делятся на открытые, с открытыми весами и закрытые. БЯМ второго типа (LLaMa 2 70B) можно самостоятельно донастраивать для решения конкретных задач, третьего – только использовать для рутинных задач (GPT-4, Claude 3, Gemini, GigaChat Pro, YandexGPT-2 и т.п.).
5. Для обучения БЯМ используются большие датасеты: мега-модели размером более 1 трлн параметров тренируются на 13 - 30 трлн токенах. Существует множество бесплатных датасетов для БЯМ с разным наполнением, но их создание уже превратилось в отдельный активно развивающийся бизнес.
6. Для оценки качества БЯМ разработано большое число бенчмарков, в том числе на базе экзаменационных вопросов для школьников, студентов и аспирантов. Рост качества БЯМ связан с повышением затрачиваемых вычислительных мощностей и прохождения кривой обучения.
7. Размер контекстного окна в БЯМ растет, что увеличивает длину памяти в диалогах чат-ботов и улучшает качество ответа на запрос пользователя.
8. Несмотря на рост оценок в тестах, БЯМ пока плохо справляются с задачами, требующими многоступенчатых рассуждений и математических действий.
9. Мультиагентные системы (МАС) – разновидность систем на базе БЯМ, в рамках которых используется несколько БЯМ, каждая из которых имеет свой ограниченный круг задач, экспертизу и доступ к инструментам решения задачи. Ожидается, что в перспективе именно мультиагентные системы будут способны решать сложные задачи, требующие ~ многоступенчатых логических суждений.
10. Хотя выход новых GPU и позволяет обучать БЯМ в разумные сроки, стоимость обучения при этом растет в разы и приближается к 1 млрд долл.
11. За 2020-2023 гг. суммарные глобальные инвестиции в GenAI составили около 50 млрд долл. (4,6 трлн руб.), из них половина пришлась на 2023 г.
12. В 2023 году ведущие разработчики БЯМ разместили рекордные заказы на H100, что обещает появление новых БЯМ с супер возможностями.

Ключевые выводы (2/4): БЯМ могут помочь российскому нефтегазу достичь технологического суверенитета, коммерческие модели для этого не пригодны

Перспективы и направления использования БЯМ в нефтегазовой отрасли

1. Нефтегазовые компании уже активно используют ИИ с глубоким обучением для работы с большими структурированными числовыми данными. При этом в отечественном нефтегазе сегодня отсутствуют промышленные решения на базе генеративных моделей. Эти технологии пока только «пилотируются» ограниченным кругом российских ВИНКов.
2. В мире ежегодно публикуется более 25 ТБ данных* в сфере НИОКР по технологиям нефтегазовой отрасли. Это базы данных по новым патентам, материалы аналитических агентств, правительственных организаций, научных и отраслевых изданий, деловых СМИ и т.д.
3. Поиск технологий для закрытия потребностей бизнеса является трудоемким и, преимущественно, ручным процессом. Чтобы быть полностью в курсе событий, эксперты должны тратить от 80% рабочего времени на изучение обновляемой отраслевой информации.
4. Использование GenAI позволяет охватить больший объем информации, сократить время на ее анализ, увеличить скорость и качество решений. Так, например, GenAI сможет повысить автоматизацию функциональных процессов, связанных с инженерией и научно-технической экспертизой. Речь прежде всего идет о прикладной экспертизе, 59% которой может быть автоматизировано. В целом потенциал автоматизации работы профессиональных групп, связанных с инженерией, при использовании нейросетей вырастет почти в два раза – до 57%.
5. Использование ИИ инструмента для поиска и обработки информации позволит решать актуальные задачи бизнеса. Потенциал БЯМ огромен: сейчас это саммаризация текста, завтра - автоматизация типовых операций, послезавтра – автономные роботы.
6. В ближайшей перспективе GenAI сможет решать относительно простые задачи: распознавание текстовых документов, поиск в базе знаний, а также обобщение документов, построение ключевых выводов. Сюда же относится бенчмаркинг, поиск аналогов; распознавание графиков, диаграмм, таблиц, формул; генерация шаблонных отчетов.
7. В классе более интеллектуальных перспективных задач – инжиниринг новых продуктов и подбор технологий, оценка состояния производственного объекта или процесса, генерация рекомендаций по предотвращению поломок оборудования, а также прикладная экспертиза для проведения ремонтов. В частности, применение GenAI позволит создавать новые технологии и продукты с помощью подхода обратного проектирования (inverse design).
8. Генеративное проектирование уже сегодня помогает сократить вес, использование материалов и, как следствие, деньги. Мультиагентные Gen AI смогут решать сложные инженерные задачи, что повысит как скорость, так и качество проектирования.
9. Суммарный отраслевой эффект от использования генеративного ИИ, может составить 343 млрд руб. в год за счет повышения производительности труда и эффективности производства. При этом наибольшее влияние (69% от этой суммы) нейросети окажут на upstream (ГРП, разработка и бурение, обустройство и капстроительство, мониторинг и управление добычей).

Ключевые выводы (3/4): Лучшие зарубежные БЯМ не отвечают на конкретные отраслевые вопросы, технология RAG улучшает качество ответов

Проблемы использования БЯМ для поиска и анализа отраслевой информации

1. Для тестирования различных БЯМ на предмет знания нефтегазовой отрасли, мы взяли конкретный запрос ПАО «Газпром нефть» по анализу применяемых технологий бурения скважин в неустойчивых аргиллитах. Для анализа были выбраны State-of-the-Art модели на март 2024: Gemini 1, Claude 3 Opus, GPT-4.
2. Результат тестирования показал, что State-of-the-art БЯМ не способны решить поставленную задачу во многом из-за отсутствия в их корпусе данных с отраслевой и страновой спецификой. Так, большая часть названий компаний, якобы использующих технологии бурения в аргиллитах, предоставленных Gemini, оказалась сгаллюцинированной, конкретика по компонентам составов отсутствует. Claude 3 не указал названия продуктов, представленные моделью компании не выпускают указанную им продукцию. GPT-4 честно написал, что у него нет данных, поэтому он представил только вымышленные названия компаний.
3. В целом оказалось, что случае анализа отраслевых данных коммерческие БЯМ дают верные ответы только на общие (не конкретные) вопросы, требующие минимальных умственных усилий и уступают человеку по глубине суждений. Для получения более корректного ответа необходимо использовать специальные техники формулирования запросов (prompt engineering), которым пользователю нужно обучаться отдельно.
4. Иностранные БЯМ строго соблюдают введенный Западом санкционный режим и ограничивают доступ к любой информации, которая может быть использована для отраслевого развития, импортозамещения и обеспечения технологического суверенитета. Кроме того, использование иностранных БЯМ с открытыми параметрами потенциально может создать угрозу IT-безопасности (внешние атаки на информационную инфраструктуру, промышленный шпионаж и т.д.).
5. Для погружения готовых БЯМ в отраслевую тематику мы дополнили их технологией RAG. Это решение было протестировано на вышеописанной отраслевой задаче и показало лучшие результаты, чем коммерческие БЯМ без RAG.
6. Мы убедились, что архитектура с RAG, реализованная в MVP ВЫГОН Консалтинг, способна решить ключевые проблемы готовых БЯМ за счет обновления данных. Для этого был реализован функционал «умного поиска» в интернете для формирования отраслевой БД, включающей статьи, отчеты и патенты.
7. В результате RAG позволил получать ответы на отраслевые вопросы на базе выдержек из реальных источников. В ответах были указаны конкретные названия технологий, ограничения «заморозки» корпуса модели были преодолены, а количество «галлюцинаций» минимизировано.

Ключевые выводы (4/4): для создания отраслевой БЯМ необходимо объединение компетенций, разделение финансирования и обмен информацией

Подходы к созданию отечественных отраслевых больших языковых моделей

1. Санкционные ограничения и высокая стоимость – основные барьеры в части роста вычислительных мощностей. По нашей оценке, число доступных для расчетов карт Tesla A100 в России составляет не более 5 тыс. шт., тогда как для создания мега-модели размером свыше 1 трлн параметров нужно более 25 тыс. карт нового поколения H100.
2. Стоимость создания «с нуля» моделей уровня 2024 года превышает 500+ млн долл. (~50+ млрд руб.). Такие расходы могут позволить себе только западные Биг Техи или консорциум из нескольких компаний.
3. Текущих суммарных инвестиций в GenAI в России достаточно для создания лишь одной мега-модели уровня 2023-2024 года, в будущем потребуется их увеличение, поскольку расходы на создание и развитие могут достигнуть 100 млрд руб.
4. Варианты создания большой отраслевой модели зависят от отношения к стоимости, скорости и качеству. На стоимость создания отраслевой БЯМ влияют три ключевых фактора. Первый – это выбор подхода, предполагающий ответ на вопрос, будет ли использована существующая модель или создана принципиально новая? Второй – это объем и качество датасетов. Третий - это размер модели.
5. Для развития GenAI в России необходимы четыре составляющие: компетенции в отрасли и ML, аппаратные мощности, инвестиции и консорциум компаний (Биг Техи, отраслевые компании, экспертные и научные организации). Проектная команда по созданию / донстройке отраслевой БЯМ должна включать не только ML-инженеров, но и отраслевых экспертов, обладающих глубокими знаниями по всем сегментам нефтегазового комплекса .
6. Консолидация компаний с целью создания отраслевой платформы GenAI позволит сократить расходы на разработку отраслевых прикладных моделей. При этом развитие отраслевых БЯМ должно происходить поэтапно, с постепенным увеличением сложности решаемых задач.
7. К созданию продвинутых мультиагентных решений на базе БЯМ, способных решать сложные отраслевые задачи (проектирование и составление концептуальных планов развития, инжиниринг новых продуктов и технологий) можно приступать только после создания более простой отраслевой БЯМ (наученной на поиск аналогов, бенчмаркинг, мониторинг технологических трендов, подготовку шаблонов отчетов).



ОБЗОР ГЛОБАЛЬНЫХ ТРЕНДОВ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ



ПЕРСПЕКТИВЫ И НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ БЯМ В НЕФТЕГАЗОВОЙ ОТРАСЛИ



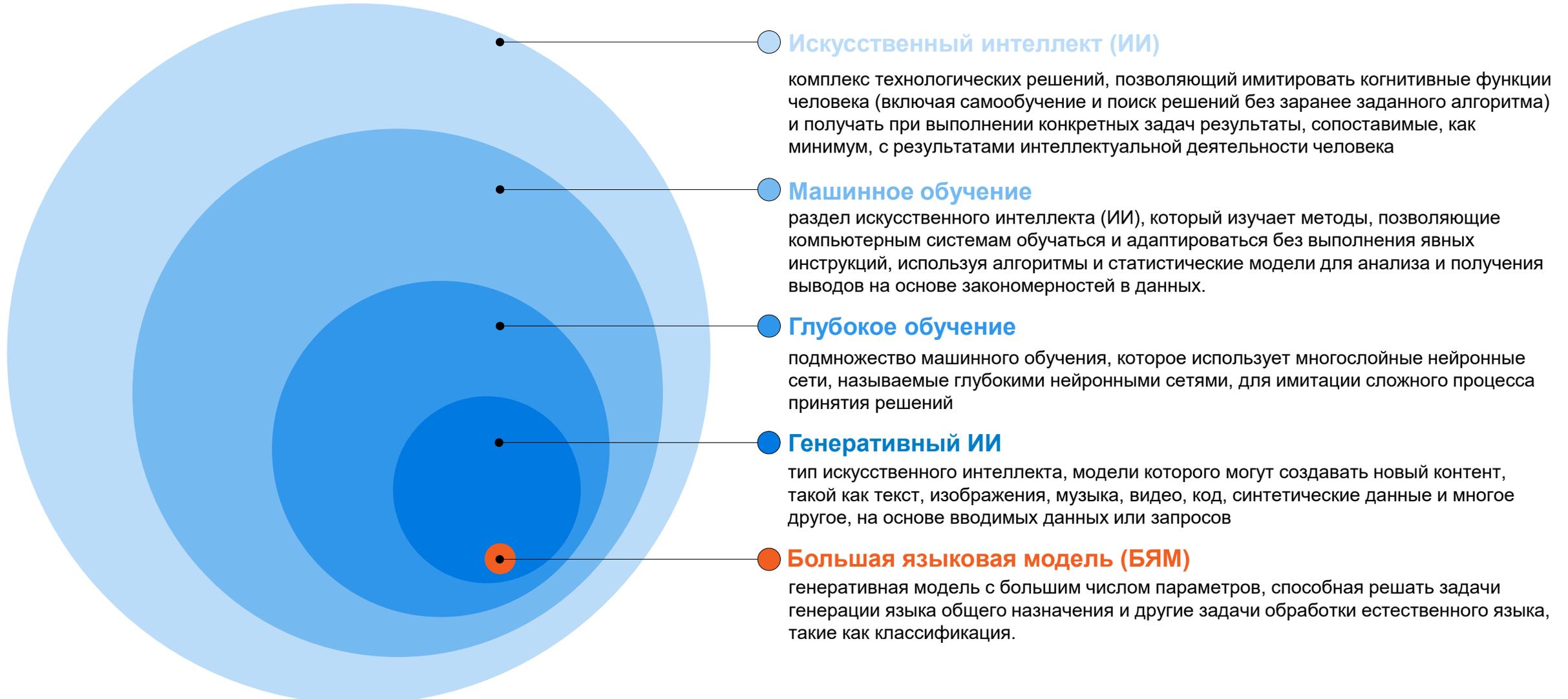
ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ БЯМ ДЛЯ ПОИСКА И АНАЛИЗА ОТРАСЛЕВОЙ ИНФОРМАЦИИ



ПОДХОДЫ К СОЗДАНИЮ ОТЕЧЕСТВЕННЫХ ОТРАСЛЕВЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Генеративный ИИ является нейросетью с глубоким обучением на неструктурированной информации, способной создавать новый контент

Определение генеративного ИИ как подкласса искусственного интеллекта с глубоким машинным обучением



В отличие от традиционного ИИ, GenAI использует глубокое обучение на большем объеме мультимодальных данных и генерирует новый контент

Сравнение генеративных моделей с традиционными нейросетями

БЯМ является частным случаем нейронных сетей, специализированным на задачах обработки естественного языка и обладающим способностью генерировать текст и изображения на основе входных данных

Характеристика	Традиционные модели	Генеративный ИИ
 Входные данные / модальность	Один тип данных для одной модели	Мультимодальные (текстовые, голосовые, физические измерения, видео и изображения)
 Архитектура	Различные архитектуры, такие как полносвязные сети, сверточные (CNN) или рекуррентные (RNN) нейронные сети	Трансформер
 Объем обучающей выборки	Размер выборки структурированных и размеченных данных: 1E9 токенов	Большая выборка неструктурированных и неразмеченных данных: 2E12
 Выходные данные	Вывод результатов обработки исследуемой выборки на основе данных по обучающей выборке	Генерация новых данных на основе корпуса и запросов пользователя
 Особенности / свойства	<ul style="list-style-type: none"> ▪ Быстрые и нетребовательные к вычислительным мощностям ▪ Работа по ранее изученным паттернам 	<ul style="list-style-type: none"> ▪ Требовательные к вычислительным мощностям ▪ Способность выявлять паттерны самостоятельно ▪ Новые свойства: саммаризация, обучение на вводимых промтах

Сегодня системы на базе GenAI требуют **широких компетенций** как со стороны **разработчиков**, так и со стороны **пользователей**

Артефакты систем на базе генеративных моделей



В процессе создания БЯМ для разных задач используются различные методы глубокого машинного обучения

Сравнение особенностей разных типов обучения БЯМ

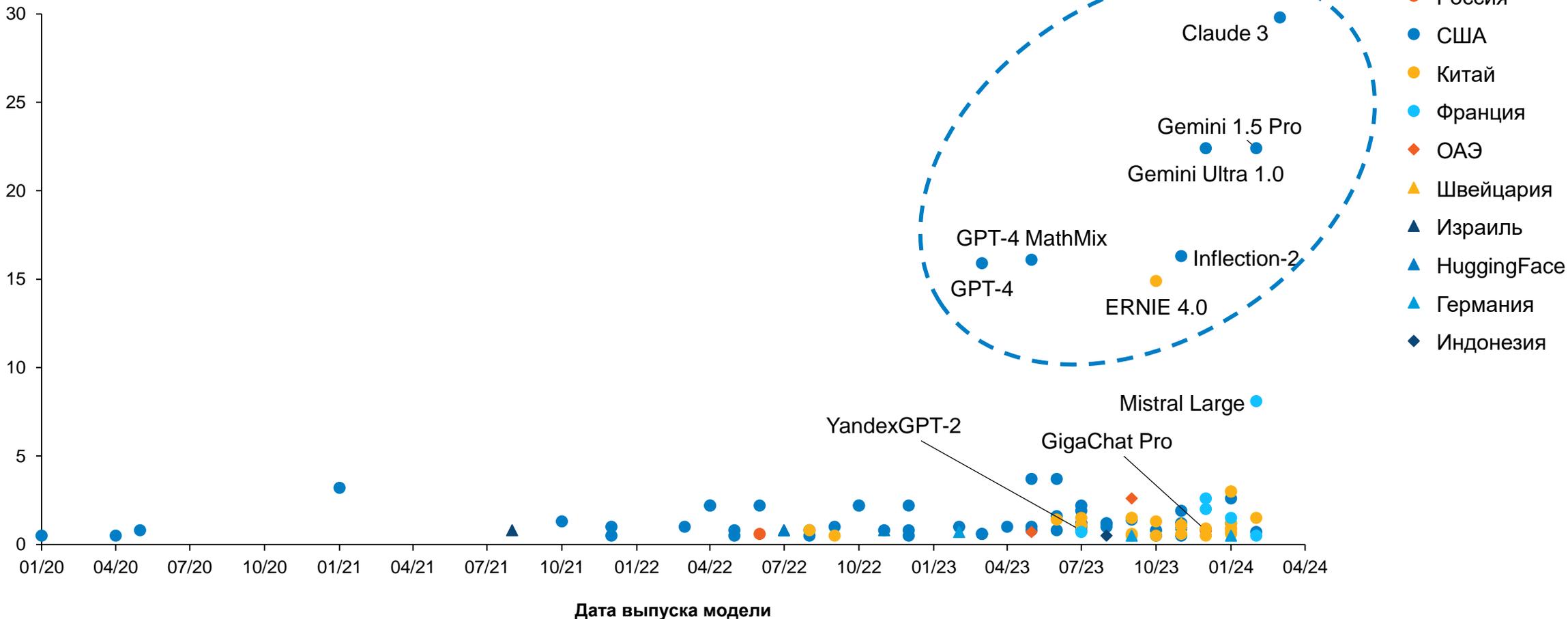
Особенности	ОБУЧЕНИЕ С УЧИТЕЛЕМ*	ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ*	ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ*
Определение	Модель обучается на размеченных данных, где для каждого примера входных данных известен желаемый выход	Модель ищет скрытые структуры в неразмеченных данных	Модель учится на основе системы наград и наказаний, стремясь к максимизации суммарного вознаграждения
Что изучается	Зависимости между входными и выходными данными	Внутренние структуры и закономерности данных	Последовательности действий для достижения цели
Типы задач	Классификация, регрессия	Кластеризация, уменьшение размерности, ассоциативные правила	Проблемы последовательного принятия решений (игры, робототехника)
Преимущества	<ul style="list-style-type: none"> Четкая оценка производительности модели Широкий спектр применения в реальных задачах 	<ul style="list-style-type: none"> Работа с неизвестными структурами данных Не требуется разметка данных, что снижает затраты на подготовку 	<ul style="list-style-type: none"> Способность приспосабливаться к изменяющимся условиям Модели могут достигать высокой степени сложности и выполнения сложных задач
Недостатки	<ul style="list-style-type: none"> Требует большого количества размеченных данных Неэффективно использовать, когда структура данных неизвестна 	<ul style="list-style-type: none"> Трудно определить корректность результатов Не подходит для задач с явными метками 	<ul style="list-style-type: none"> Требует большого количества времени и данных для обучения Сложность в настройке и интерпретации
Примеры использования	Распознавание речи, обработка изображений, прогнозирование финансовых трендов	Анализ социальных сетей, сегментация рынка, рекомендательные системы	Автономные транспортные средства, игровые боты, роботы для производственных систем

* Supervised learning, Unsupervised learning, Reinforcement learning

Каждый месяц в мире появляются 8-10 новых моделей, количество мега-моделей растёт, лидируют американские команды OpenAI, Google, Anthropic, Mistral, Meta*

Эволюция мощности моделей

ALScore **



* признана экстремистской и запрещена в РФ

** ALScore - быстрая и грубая метрика мощности модели. Формула расчета: $ALScore = \frac{\sqrt{\text{Параметры} \times \text{Токены}}}{300}$

Источник: [David McCandless, Tom Evans, Paul Barton](#), 2024 LifeArchitect.ai data, ВЫГОН Консалтинг

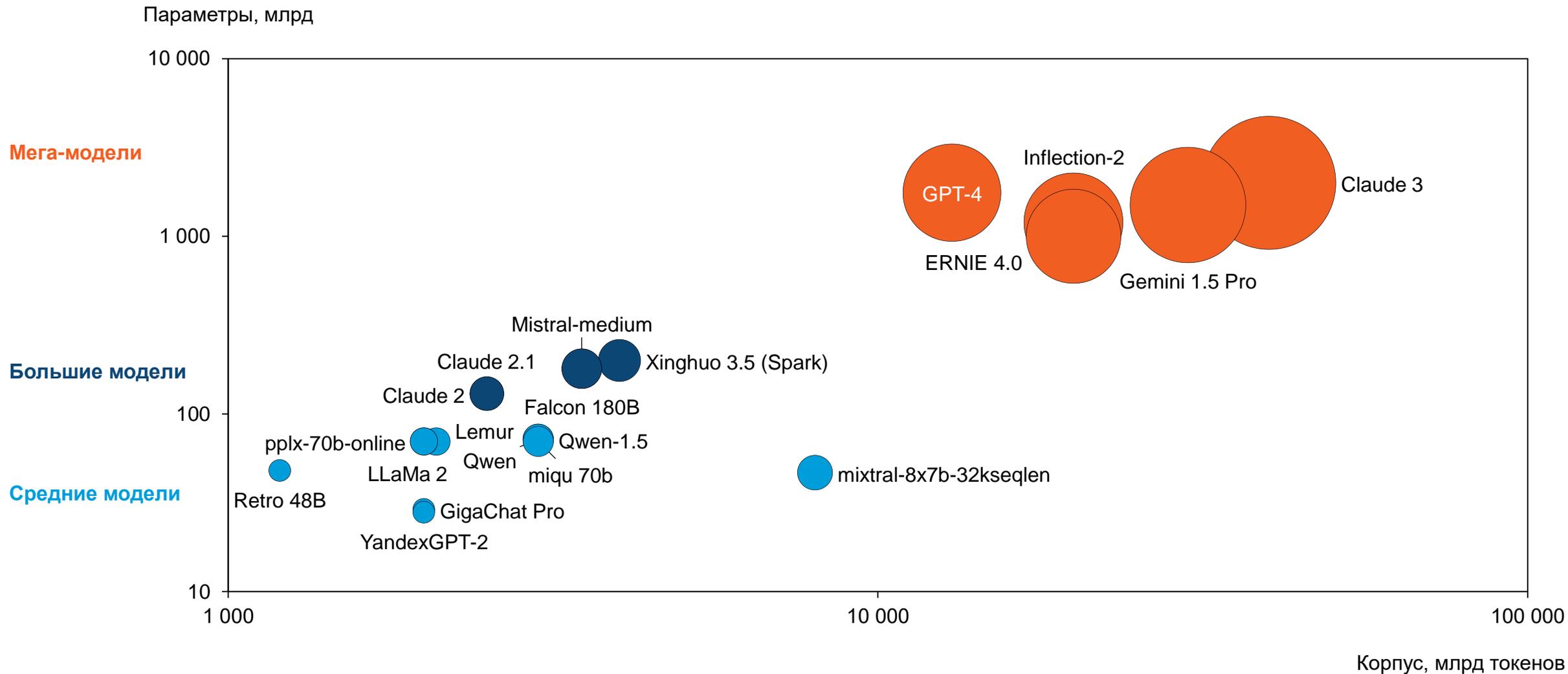
По степени открытости большие языковые модели делятся на открытые, с открытыми весами и закрытые

Сравнение БЯМ по степени открытости

Свойство	Описание	Открытые	Открытые веса	Закрытые
Вывод (Inference)	Генерация выходных данных с помощью предварительно обученной (pre-trained) модели	✓	✓	✓
Донастройка (Fine-tuning)	Возможность дополнительного обучения pre-trained модели для решения определенных задач	✓	✓	✓
Доступность весов	Весы – это большой массив (1-1000 млрд) параметров (чисел с плавающей точкой) нейросети, которые вычисляются в результате ее предварительного обучения, и не меняются при генерации. Могут быть изменены в процессе дообучения.	✓	✓	✗
Приватность	Возможность запуска модели на собственной ИТ инфраструктуре	✓	✓	✗
Архитектура	Детальное описание архитектуры модели, формул, алгоритмов, методов, значений гиперпараметров, используемых для ее предварительного обучения, позволяющее воспроизвести модель «с нуля»	✓	✗	✗
Данные для предварительного обучения	Описание используемых данных для предварительного обучения, способы их получения, предобработки. Данные могут включать тексты из интернета, предварительно собранные открытые и закрытые наборы данных (datasets), книги, статьи, отчеты, исходный код, синтетические данные, полученные с помощью других моделей.	✓	✗	✗
Примеры моделей		OPT, MPT	LLaMa, Mistral, Gemma, Kandinsky	GPT, Claude, Gemini

Для обучения БЯМ используются большие датасеты, мега-модели размером более 1 трлн параметров тренируются на 13 - 30 трлн токенах

Зависимость размера модели и обучающего датасета

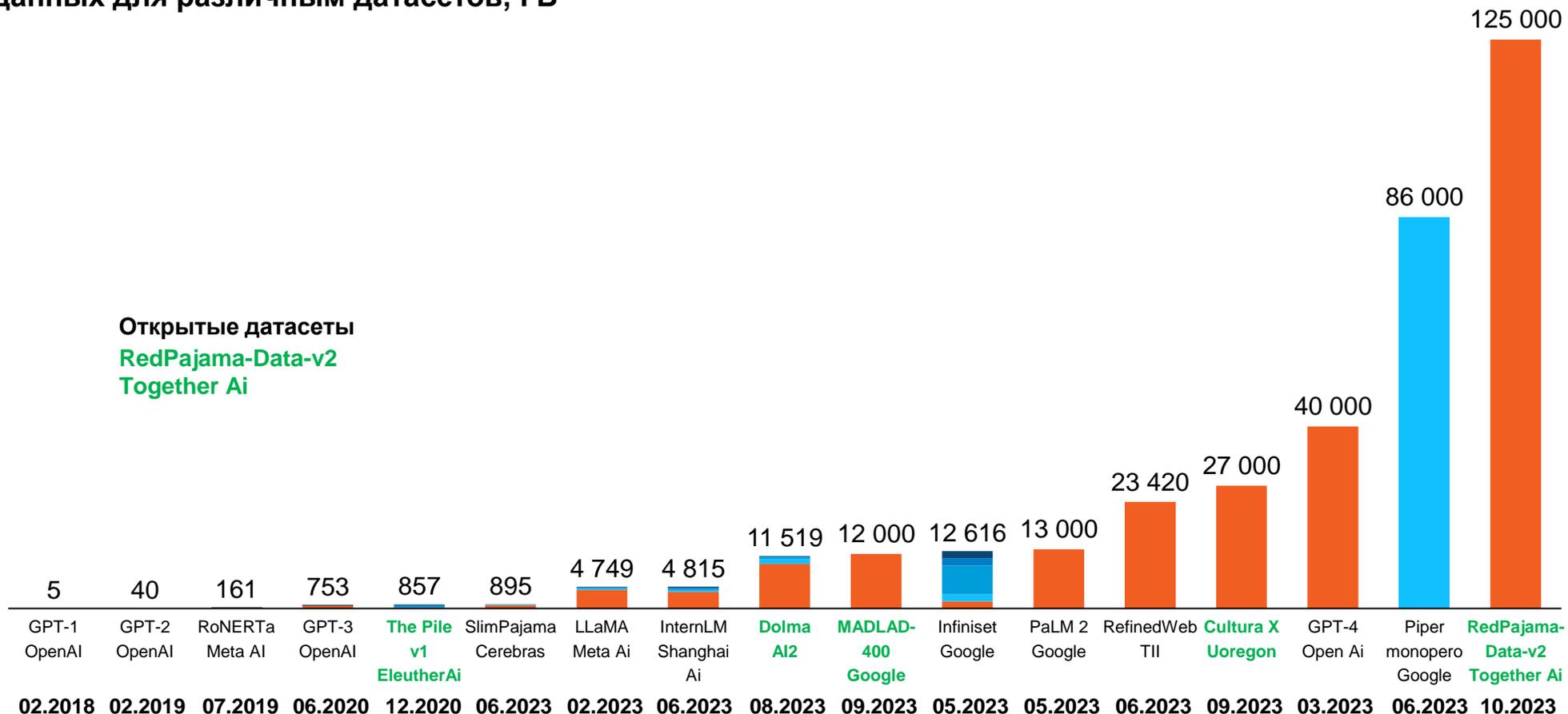


Существует множество бесплатных датасетов для БЯМ с разным наполнением, но их создание уже превратилось в отдельный активно развивающийся бизнес

Соотношения ресурсов данных для различным датасетов, ГБ

Данные для датасетов

- Википедия:**
многоязычная интернет энциклопедия
- Книги/Журналы:**
Google Books, arXiv, Project Gutenberg, HathiTrust, DOAJ
- Форумы:**
Reddit, Twitter, Discord, Stack Exchange
- Код:**
GitHub, Stack Overflow, Hacker News
- Веб-сайты**



Google Books, arXiv, Project Gutenberg, HathiTrust и DOAJ – масштабные онлайн-ресурсы, предоставляющие доступ к миллионам оцифрованных книг, научных статей, препринтов и других документов из различных областей знаний

GitHub, Stack Overflow и Hacker News – популярные онлайн-платформы, ориентированные на разработчиков программного обеспечения и технологическое сообщество, предоставляющие возможности для совместной работы над кодом, хостинга проектов, задавания и ответов на вопросы по программированию, а также обсуждения новостей, стартапов и тенденций в сфере технологий

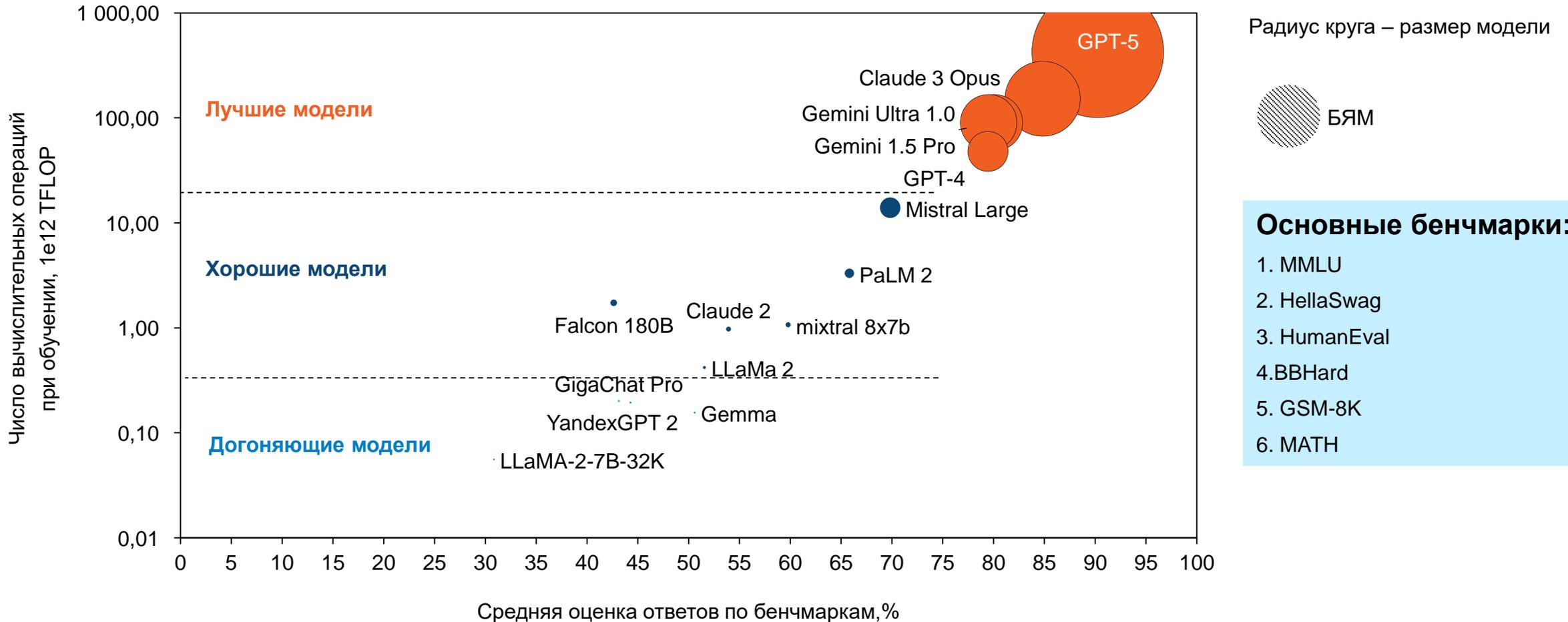
Для оценки качества БЯМ разработано большое число бенчмарков на базе экзаменационных вопросов для школьников, студентов и аспирантов

Описание тестов, используемых для сравнения качества БЯМ

Метод	Описание
MMLU (Massive Multitask Language Understanding)	Бенчмарк, который позволяет тестировать качество ответов БЯМ по 57 темам, охватывающим естественные, социальные и гуманитарные науки
GPQA (Graduate Program Question Answering)	Бенчмарк для оценки возможностей БЯМ отвечать на вопросы, особенно требующие сложных многошаговых рассуждений на основе текстов
ARC – Reasoning benchmark	Бенчмарк, который позволяет оценить способности БЯМ к рассуждению и пониманию . Набор данных содержит около 7800 научных вопросов для учащихся 3-9 классов по темам биологии, физики, химии и др. Для ответа на вопросы требуется логическое мышление и способность делать выводы , а не просто набор знаний
GSM8K (Grade School Math 8K)	Бенчмарк, состоящий из 8500 высококачественных лингвистически разнообразных школьных задач по математике. Решение этих задач занимает от 2 до 8 шагов, а решения в основном включают выполнение последовательности элементарных вычислений с использованием основных арифметических операций для получения окончательного ответа
MGSM (Multilingual Grade School Math)	Набор задач на различных языках (250 задач из GSM8K), который позволяет оценить качество ответов БЯМ на уровне учащихся начальной школы
HumanEval (Code)	Бенчмарк, позволяющий оценить способность БЯМ генерировать фрагменты кода на основе описаний на естественном языке, оценивая его возможности «кодинга»
DROP (Discrete Reasoning Over Paragraphs)	Бенчмарк, в котором БЯМ должны извлекать релевантную информацию из фрагментов текста , прежде чем выполнять над ними дискретные шаги рассуждения (например, сортировать или подсчитывать элементы), чтобы получить правильный ответ
HellaSwag	Бенчмарк, который оценивает способности БЯМ к критическому мышлению , требуя от него выбора наиболее правдоподобного продолжения заданной ситуации

Рост качества БЯМ связан с повышением затрачиваемых вычислительных мощностей: больше мощностей → выше сложность → лучше модель

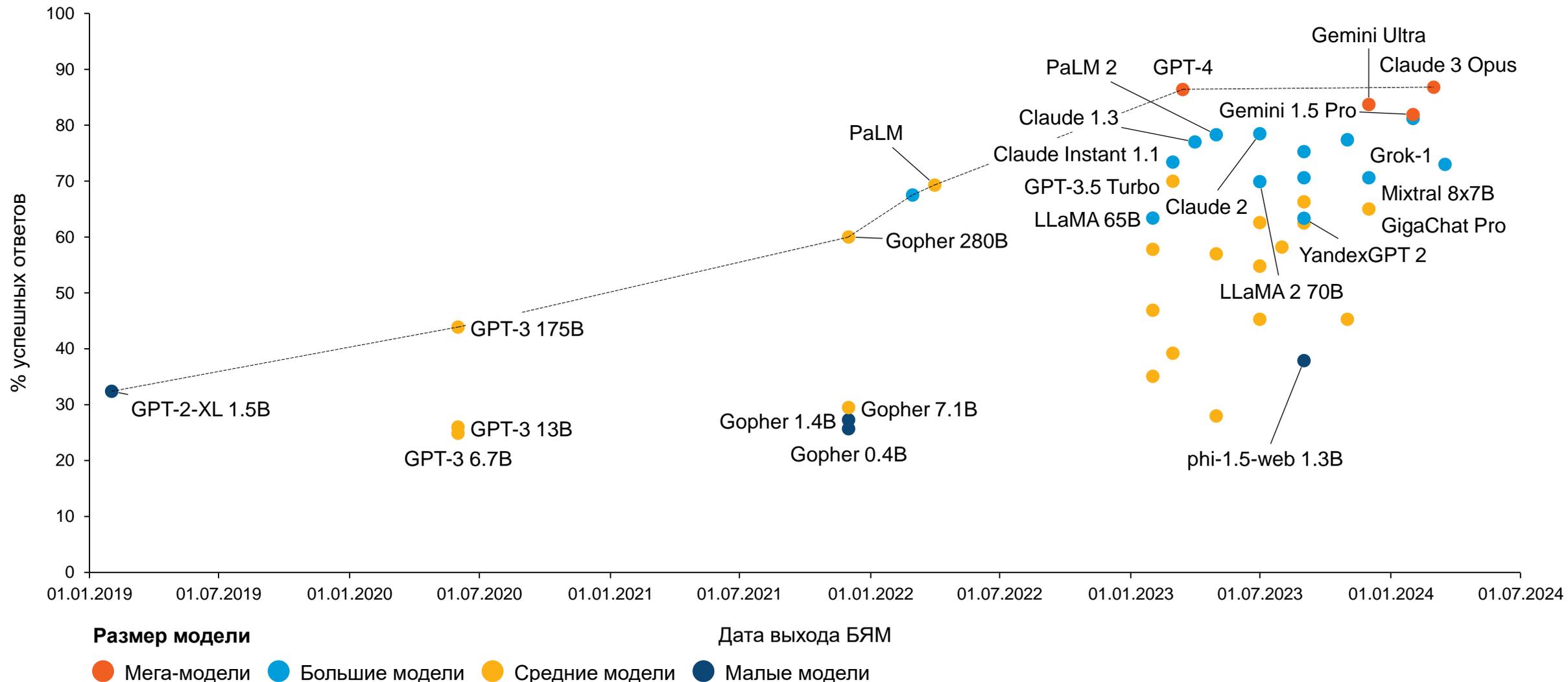
Оценка* качества модели в зависимости от их сложности



* По GPT-5, Yandex GPT-2 приведены оценки ВЫГОН Консалтинг

Улучшение качества моделей происходит не только за счет роста их размера, но и прохождения кривой обучения

Сравнение БЯМ по MMLU*

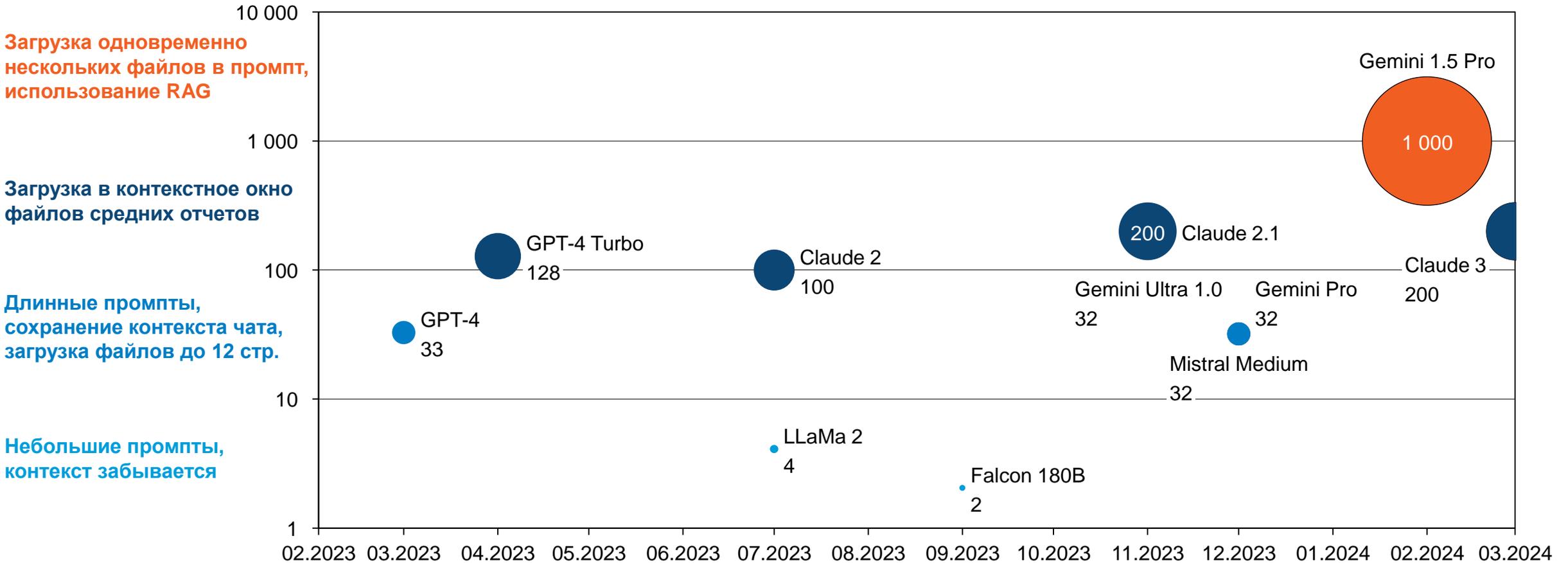


* Massive Multitask Language Understanding – сборник экзаменационных вопросов по различным областям знаний, разработанный для тестирования нейросетей ведущими университетами США
 Источник: [PapersWithCode](#), [2024 LifeArchitect.ai data](#), ВЫГОН Консалтинг

Размер контекстного окна в БЯМ растет, что увеличивает длину памяти в диалогах чат-ботов и улучшает качество ответа на запрос пользователя

Динамика размера контекстного окна в БЯМ, тыс. токенов

Контекстное окно – это ограничение на количество токенов, которые модель может анализировать и использовать для понимания контекста и генерации текста. Размер контекстного окна определяет, насколько далеко модель может "вспомнить" информацию в тексте при генерации ответов. Если текст, который нужно обработать, превышает размер контекстного окна, модель не сможет учесть весь текст одновременно. В таком случае, информация за пределами контекстного окна не будет учтена при генерации ответа.



Несмотря на рост оценок в тестах, БЯМ пока плохо справляются с задачами, требующими многоступенчатых рассуждений и математических действий

Кейс на рассуждения (reasoning) для различных БЯМ

Решение проблем – способность больших языковых моделей анализировать, интерпретировать и отвечать на вопросы или задачи, требующие логического мышления, понимания контекста и применения знаний. Это включает в себя идентификацию вопроса или проблемы, поиск релевантных знаний или информации, применение правил или концепций для построения решения и представление этого решения в понятном виде.

Пример задачи (GS8MK)

Медведь готовится к зимней спячке, и ему необходимо набрать 1000 фунтов. В конце лета медведь питается ягодами и мелкими лесными животными. Осенью он питается желудями и лососем. Летом он набирал пятую часть необходимого веса за счет ягод, а осенью - вдвое больше за счет желудей. Лосось составил половину оставшегося веса, который ему нужно было набрать. Сколько килограммов он набрал, питаясь мелкими животными? Приведи решение по шагам. Ответ укажи в фунтах.

Claude 3 Sonnet

Дано:

- Общий необходимый вес для медведя - 1000 фунтов.
- Летом он набрал 1/5 от необходимого веса за счет ягод.
- Осенью он набрал вдвое больше, чем летом, за счет желудей.
- Лосось составил половину оставшегося веса, который ему нужно было набрать.

Шаг 1: Рассчитаем вес, набранный за счет ягод.

Вес от ягод = $1/5 \times 1000$ фунтов = 200 фунтов

Шаг 2: Рассчитаем вес, набранный за счет желудей.

Вес от желудей = $2 \times$ Вес от ягод = 2×200 фунтов = 400 фунтов

Шаг 3: Рассчитаем оставшийся вес, который нужно было набрать после ягод и желудей.

Оставшийся вес = Общий необходимый вес - Вес от ягод - Вес от желудей

= 1000 фунтов - 200 фунтов - 400 фунтов

= 400 фунтов

Шаг 4: Рассчитаем вес, набранный за счет лосося.

Вес от лосося = $1/2 \times$ Оставшийся вес = $1/2 \times 400$ фунтов = 200 фунтов

Шаг 5: Рассчитаем вес, набранный за счет мелких лесных животных.

Вес от мелких животных = Общий необходимый вес - Вес от ягод - Вес от желудей - Вес от лосося

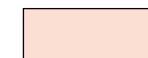
= 1000 фунтов - 200 фунтов - 400 фунтов - 200 фунтов

= 200 фунтов

Модель	Ответ
Claude 3	200
GPT-4	200
LLaMa 2 70B	200
GigaChat	0
LLaMa 2 13B	600
LLaMa 2 7B	800
Mistral 7B	200
Yandex GPT-2	250



Правильный ответ

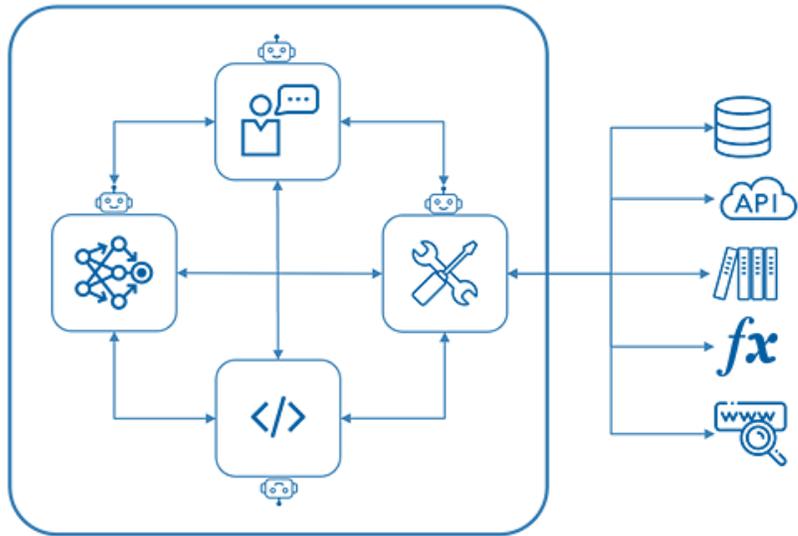


Неправильный ответ

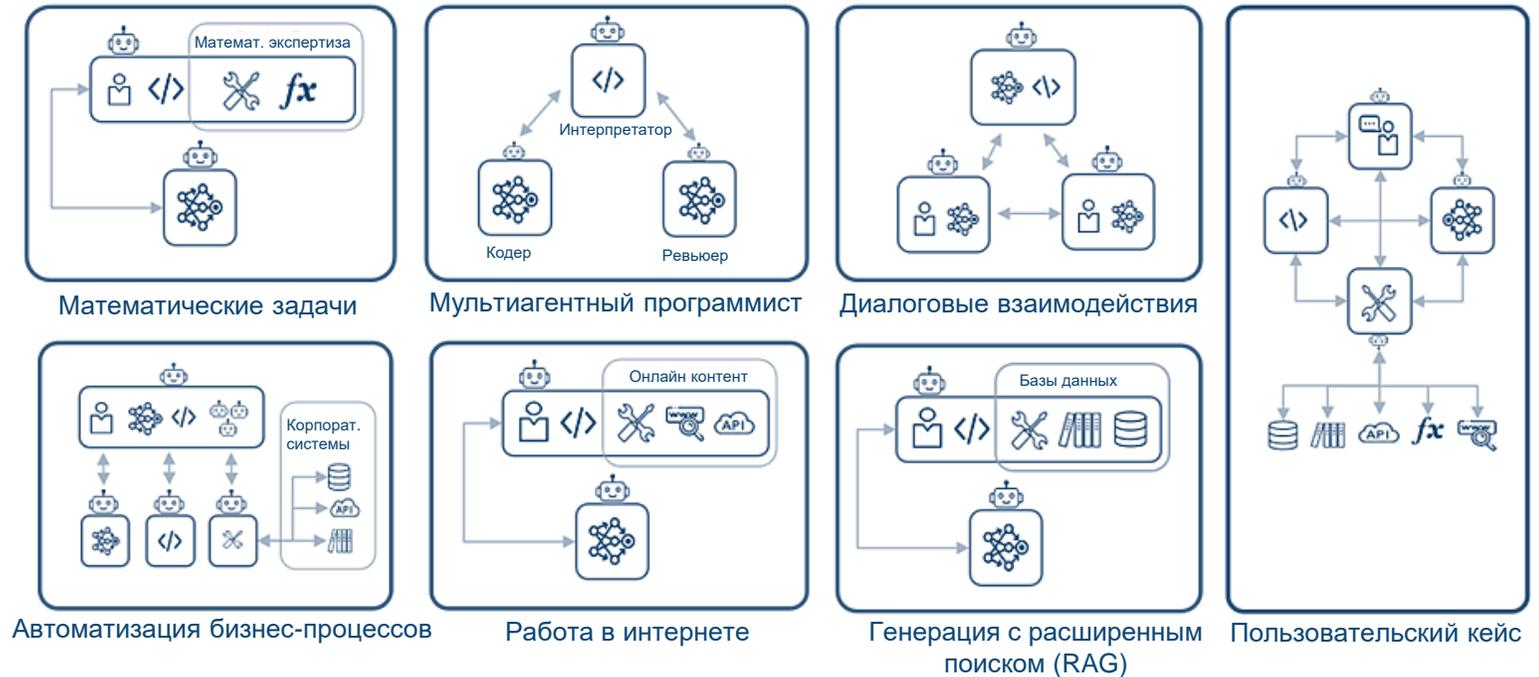
Ожидается, что в перспективе мультиагентные системы будут способны решать сложные задачи, требующие многоступенчатых логических суждений

Мультиагентные системы (МАС) – разновидность систем на базе БЯМ, в рамках которых используется различные модели, каждая из которых хорошо решает свой ограниченный круг задач, имеет экспертизу и доступ к внешним инструментам

Примерная схема мультиагентной системы



Перечень задач, которые способны решать мультиагентные системы



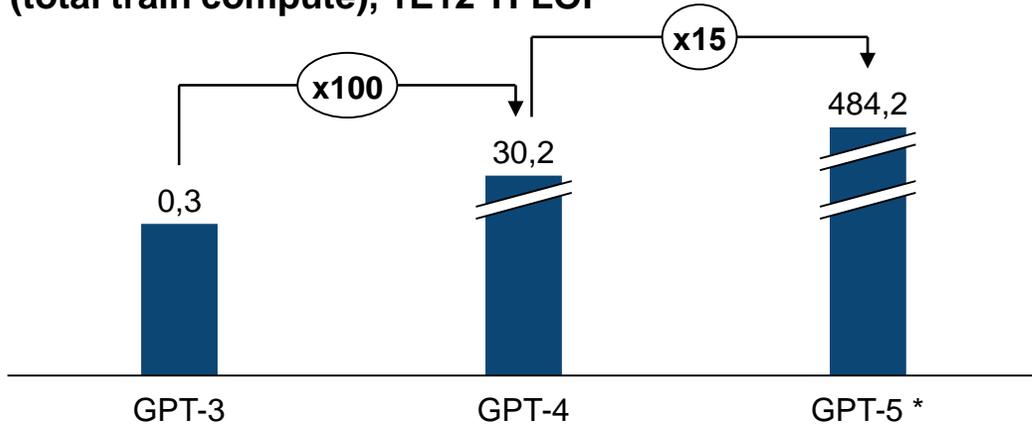
Условные обозначения

- Генеративная модель
- Агент с доступом к набору инструментов
- Доступ к приложениям через API
- Поиск в интернете

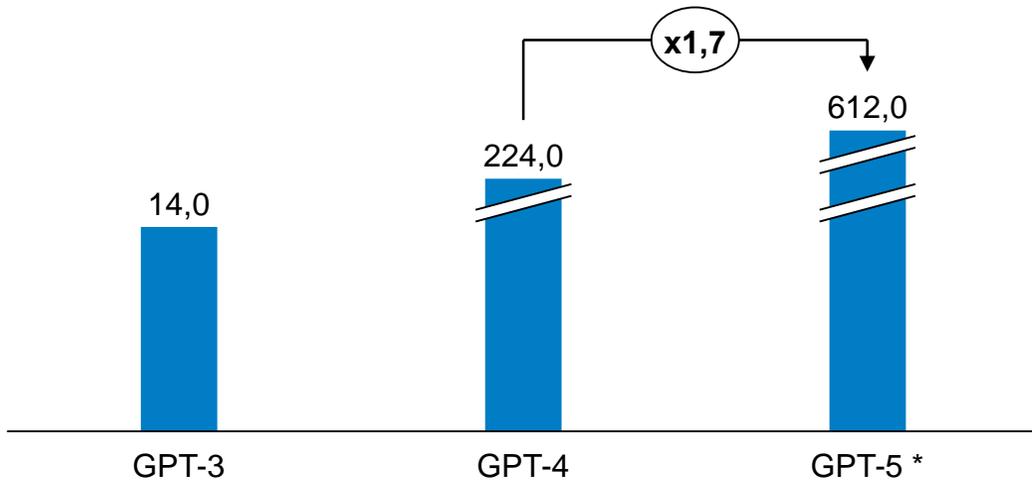
- Агент кодер / Интерпретатор
- Языковой агент
- Корпоративные Базы знаний
- Корпоративные БД
- ПО для математики

С каждым годом для **обучения БЯМ** требуется использование всё **бОльших вычислительных мощностей**, **стоимость обучения приближается к 1 млрд долл.**

Число вычислительных операций при обучении (total train compute), 1E12 TFLOP



Оценка стоимости обучения**, млн долл.



Основные параметры графических чипов

Название	Год релиза	Поколение	Объем памяти, Гб	Теоретическая производительность FP16, TFLOP/s	Стоимость покупки, долл.	Стоимость аренды, долл./ час
Tesla P100	2016	Pascal	16	19,0	200	н/д
Tesla V100	2017	Volta	32	31,3	6 000	2,5
Tesla A100	2020	Ampere	80	78,0	15 000	3,9
Tesla H100	2023	Hopper	96	248,3	30 000	4,2
GeForce RTX 4090	2023	GeForce 40	24	93,2	1 700	н/д



NVIDIA GeForce RTX 4090 – последнее поколение потребительской видеокарты, предназначенная для установки в игровые ПК

Tesla – поколения серверных видеокарт, предназначенных для решения задач машинного обучения



Для обучения GPT-4 использовали **25 тыс. карт A100** на протяжении **95 дней**

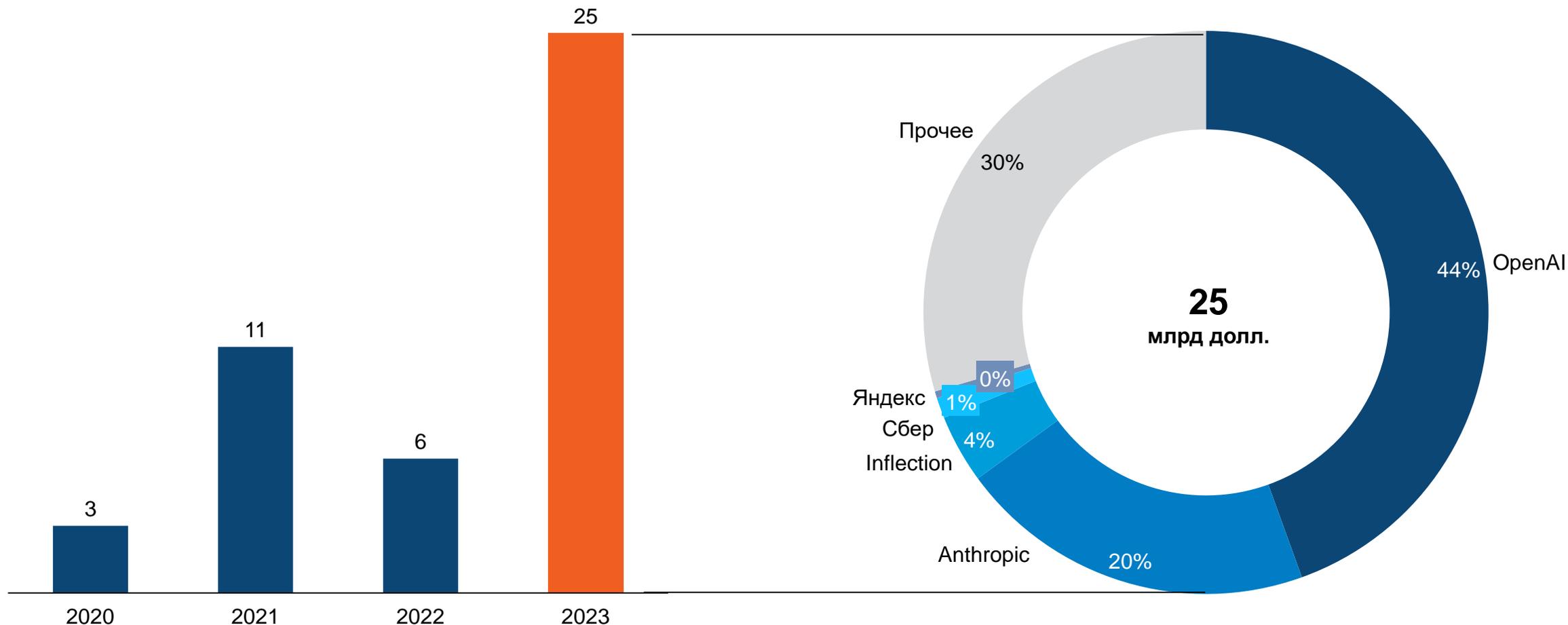
Предполагается, что Для GPT-5 использовано **50 тыс. карт H100** на протяжении **120 дней**

* Оценка ВЫГОН Консалтинг ** Стоимость аренды GPU на обучение, без учета расходов на программистов

Только за 2020-2023 гг. суммарные глобальные инвестиции в GenAI составили около 50 млрд долл., из них половина пришлась на 2023 г.

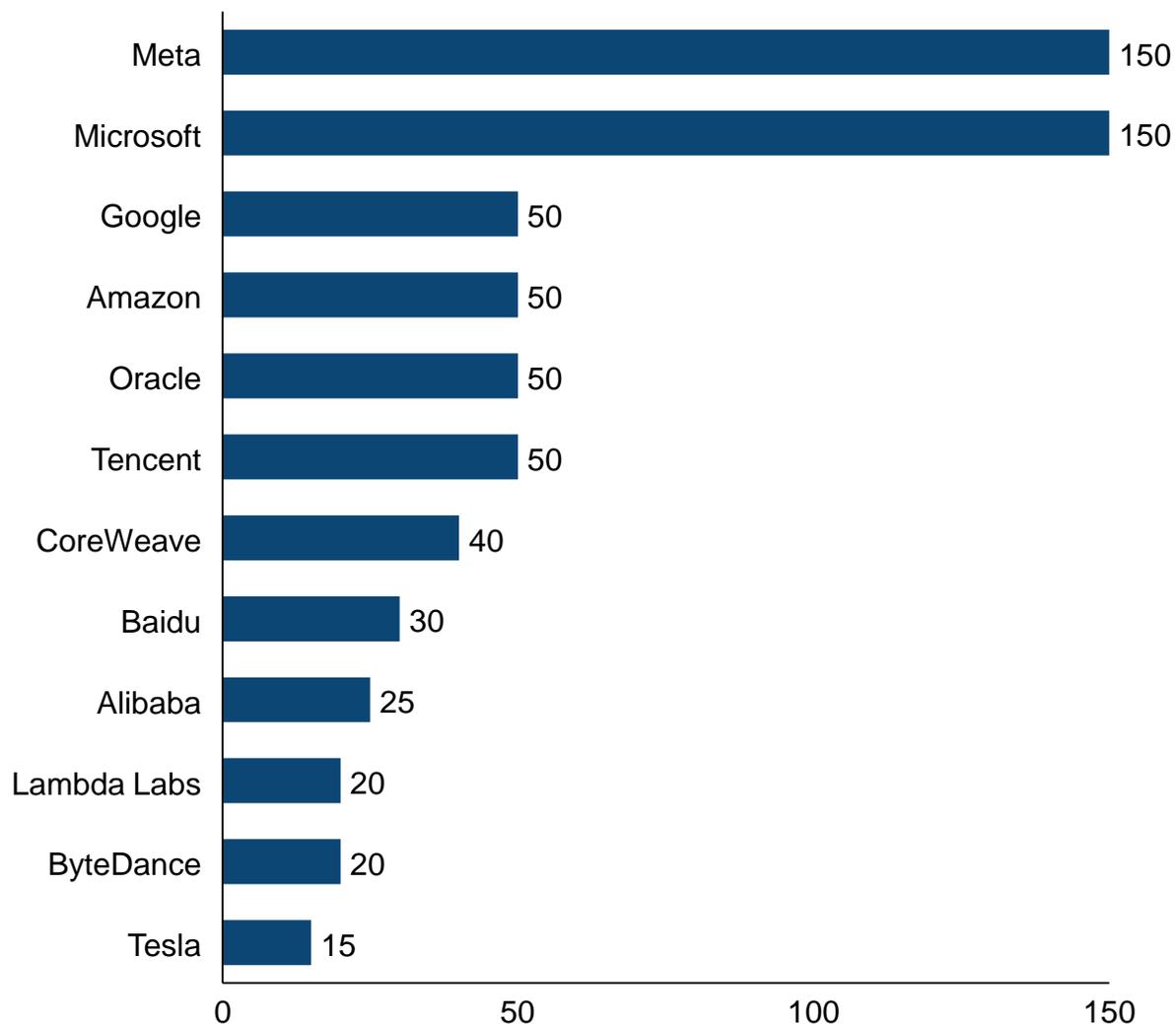
Динамика инвестиций в GenAI 2020-2023, млрд долл.

Инвестиций в GenAI в 2023 году, %

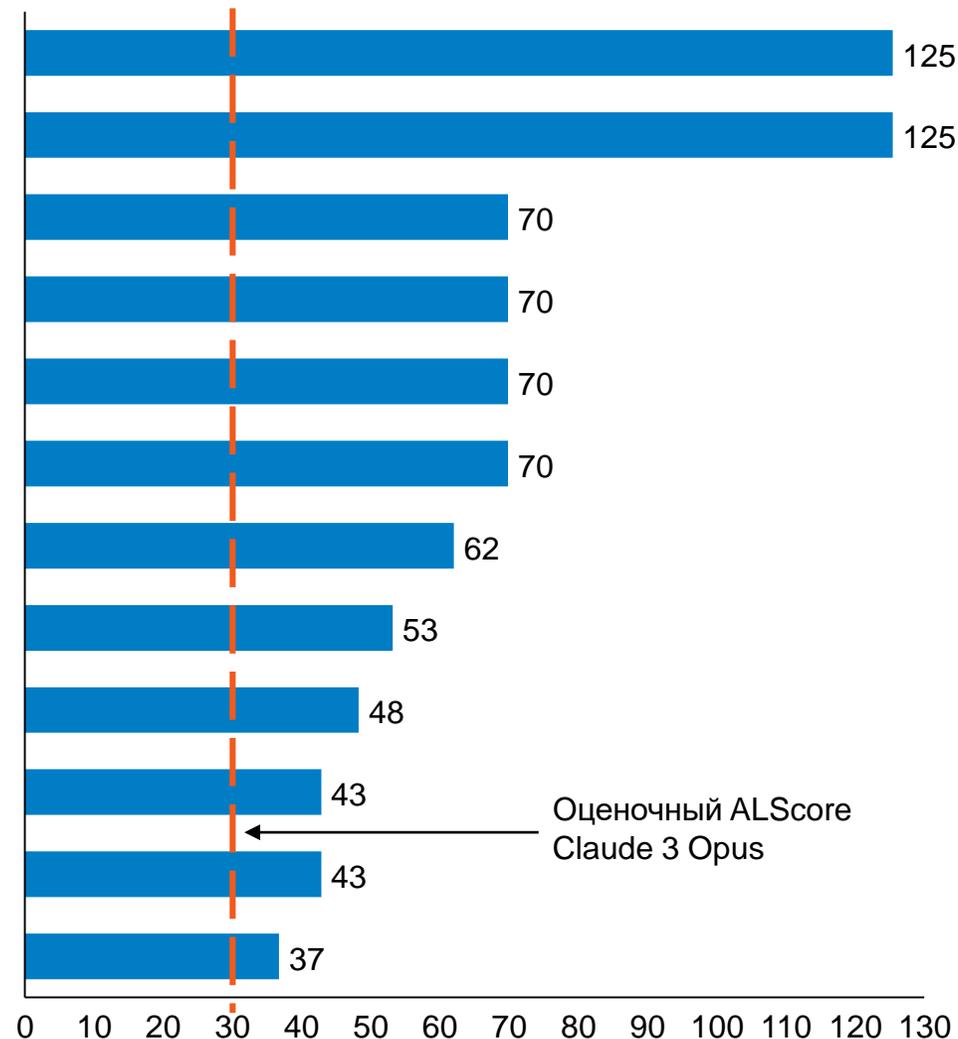


В 2023 году ведущие разработчики БЯМ разместили рекордные заказы на чипы H100, что обещает появление новых БЯМ в 2024 с супер возможностями

Заказы крупных компаний на графические процессоры H100 в 2023 году, тыс. штук



Оценочный ALScore моделей, которые могут быть натренированы на заказанных GPU H100 за 120 дней





ОБЗОР ГЛОБАЛЬНЫХ ТРЕНДОВ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ



ПЕРСПЕКТИВЫ И НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ БЯМ В НЕФТЕГАЗОВОЙ ОТРАСЛИ



ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ БЯМ ДЛЯ ПОИСКА И АНАЛИЗА ОТРАСЛЕВОЙ ИНФОРМАЦИИ



ПОДХОДЫ К СОЗДАНИЮ ОТЕЧЕСТВЕННЫХ ОТРАСЛЕВЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Нефтегазовые компании уже активно используют ИИ с глубоким обучением для работы с большими структурированными числовыми данными

Примеры использования ИИ с глубоким обучением в основных процессах нефтегазовой отрасли

Технология	Входные данные	Цель
 <p>Когнитивный геолог</p> <p>Цифровая скважина</p> <p>ИИ-интерпретатор сеймики</p> <p>Алхимик</p>	<p>Цифровая геологическая информация (скважинные данные, сеймика и т.д.)</p> <p>Числовые показатели с буровой</p> <p>Сейсмические сигналы</p> <p>Физико-химические показатели (вязкость, плотность и т.д.)</p>	<p>Трансформация подхода к анализу неопределенностей и построения концептуальных геологических моделей</p> <p>Контроль технологического процесса бурения скважин, внесение корректировок в режиме реального времени</p> <p>Обработка больших объемов цифровых сейсморазведочных данных с помощью предобученного ИИ</p> <p>Создание многокомпонентных рецептур моторных масел с использованием ИИ</p>
 <p>РН-нейросети</p>	Цифровая геологическая и технологическая информация	Самостоятельная оптимизация вариантов размещения новых скважин, параметров проведения гидроразрыва пласта и т.д.
 <p>ТН-Нейрограм</p>	Числовые скважинные данные	Подбор скважин для ГТМ и расчёт гидроразрыва пластов
 <p>СОУиКА</p>	Цифровые данные термо-, виброакустических датчиков	Определение потенциальных опасностей на трубопроводе
 <p>Предиктивное обслуживание</p>	Данные с датчиков оборудования	Снизить количество незапланированных простоев на морских нефтяных и газовых платформах
 <p>MobilServ</p> <p>Интерпретация сейсмических данных</p>	<p>Информация об оборудовании, образце, расходе материалов</p> <p>Сейсмические данные</p>	<p>Сократить время плановых простоев, расход смазочных материалов и затраты.</p> <p>Систематизировать и сократить во времени процесс сбора сейсмических данных, разрозненных в сотнях и тысячах приложений</p>
 <p>Shell Corrosion Advanced Risk Modelling and Analytics</p>	Информации с датчиков производственных объектов	Минимизация рисков возникновения коррозии и снижение необходимости полевых проверок

В мире ежегодно публикуется более 25 ТБ данных* в сфере НИОКР по технологиям нефтегазовой отрасли

Источники информации в сфере ТЭК (преимущественно нефти и газа)

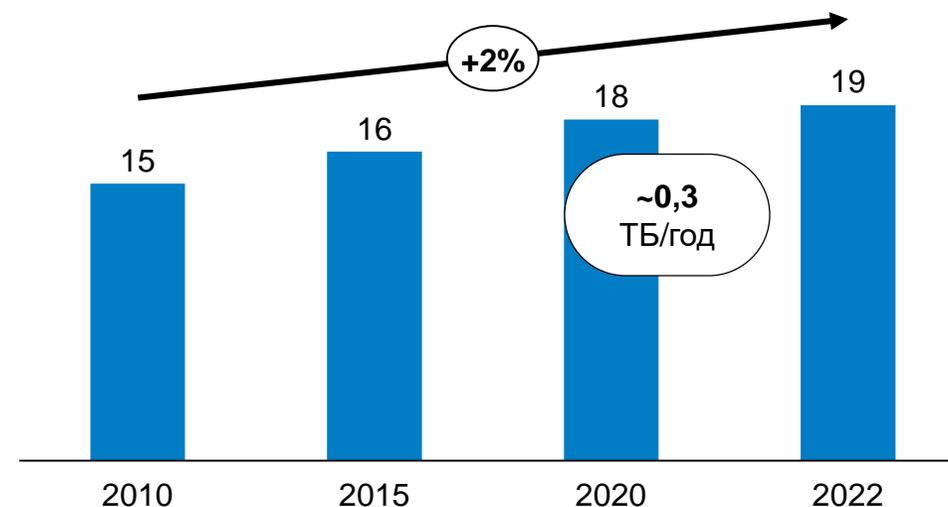
Научные статьи в сфере инжиниринга

Кол-во опубликованных статей по тематике инжиниринга, тыс.



Патенты в сфере инжиниринга

Кол-во патентов в сфере нефтегаза, тыс.



Прочие источники

- Профильные СМИ (от 1 до 20 новостей в день)
- Нефтегазовые журналы (в среднем 1 статья в день)
- Публикации компаний-производителей (от 1 до 10 постов в месяц)
- Прочие

~5 ТБ/год

- Более **2 млн научных статей** в сфере нефтегаза, более **15 млн** в смежных областях (инжиниринг) за 20-21 века
- Более **150 тыс. патентов** в сфере нефтегаза
- Прочие

* Исходных данных только в статьях и патентах

Поиск технологий для закрытия потребностей бизнеса является трудоемким и, преимущественно, ручным процессом

Существующая схема алгоритма поиска технологического решения



На поиск одной технологии может уйти **несколько месяцев полной занятости специалиста!**



Закрытие потребностей бизнеса может быть срочным. В этом случае трата времени на каждом этапе вызывает альтернативные издержки в размере потраченного на поиски время (дни) * прибыль (руб./день)



«Ручной» алгоритм не гарантирует нахождение оптимального решения, оно могло не попасть в начальный список источников

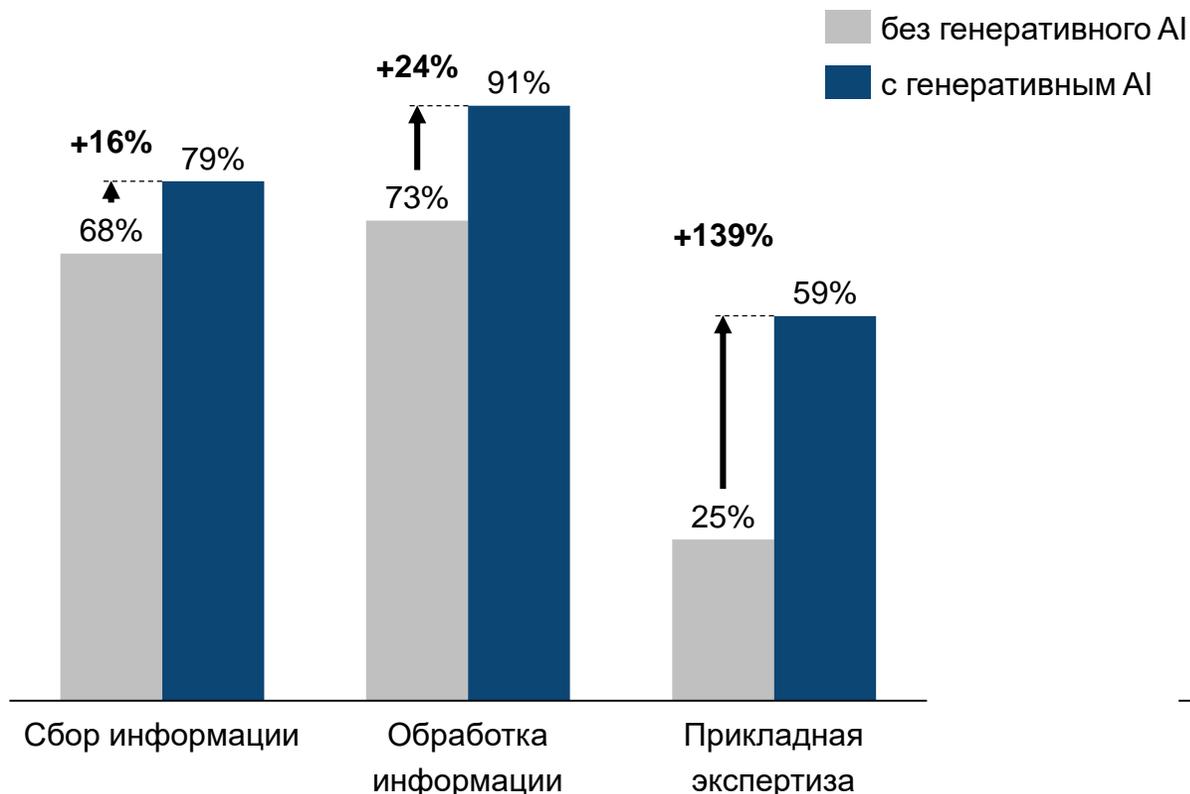
Чтобы быть в курсе событий, эксперты должны тратить от 80% рабочего времени на изучение обновляемой отраслевой информации

Примеры материалов, которые ежегодно обязан изучать сотрудник аналитического подразделения

Источник	Примеры	Объем информации	
		МБ	Час.
Мировые агентства		600	500
Глобальные нефтегазовые компании		50	50
Российские компании (годовые, финансовые, ESG отчеты и презентации)		245	150
Консалтинговые агентства и исследовательские центры		600	500
Банки		90	75
ФОИВ		50	30
Отраслевые журналы		300	250
Прочее (обновления НПА, СМИ и т.д)		150	100
Итого		2085	1655

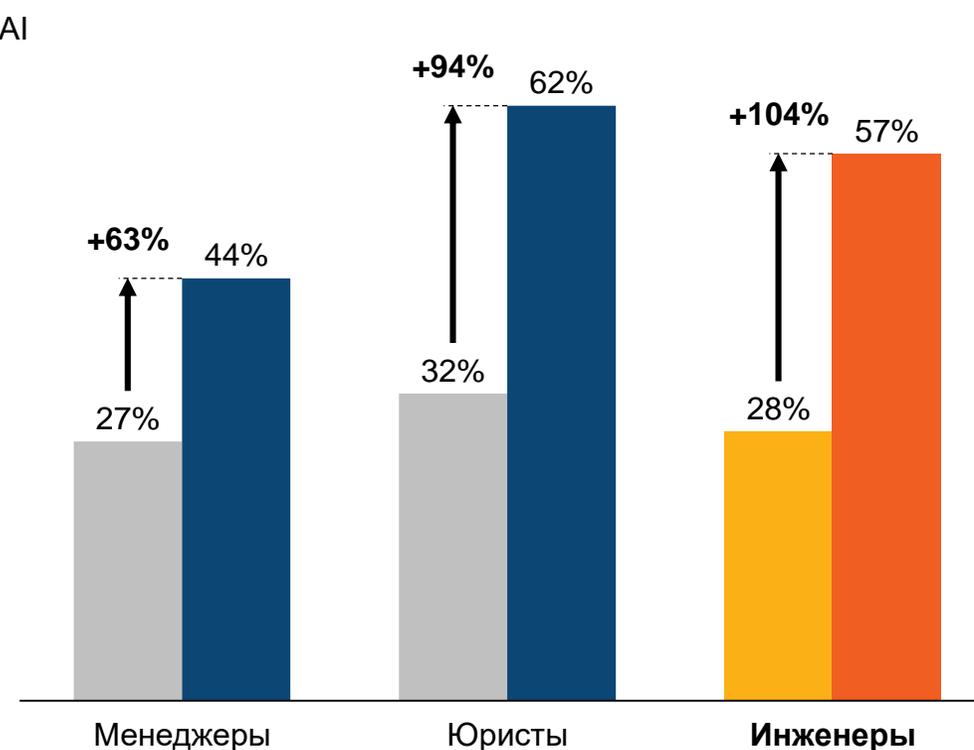
Использование GenAI позволяет охватить больший объем информации в сфере инженерии и сократить время на ее анализ

Потенциал автоматизации функциональных процессов, связанных с инженерией



GenAI может оказать наибольшее влияние на прикладную экспертизу, которая ранее практически не имела потенциала для автоматизации

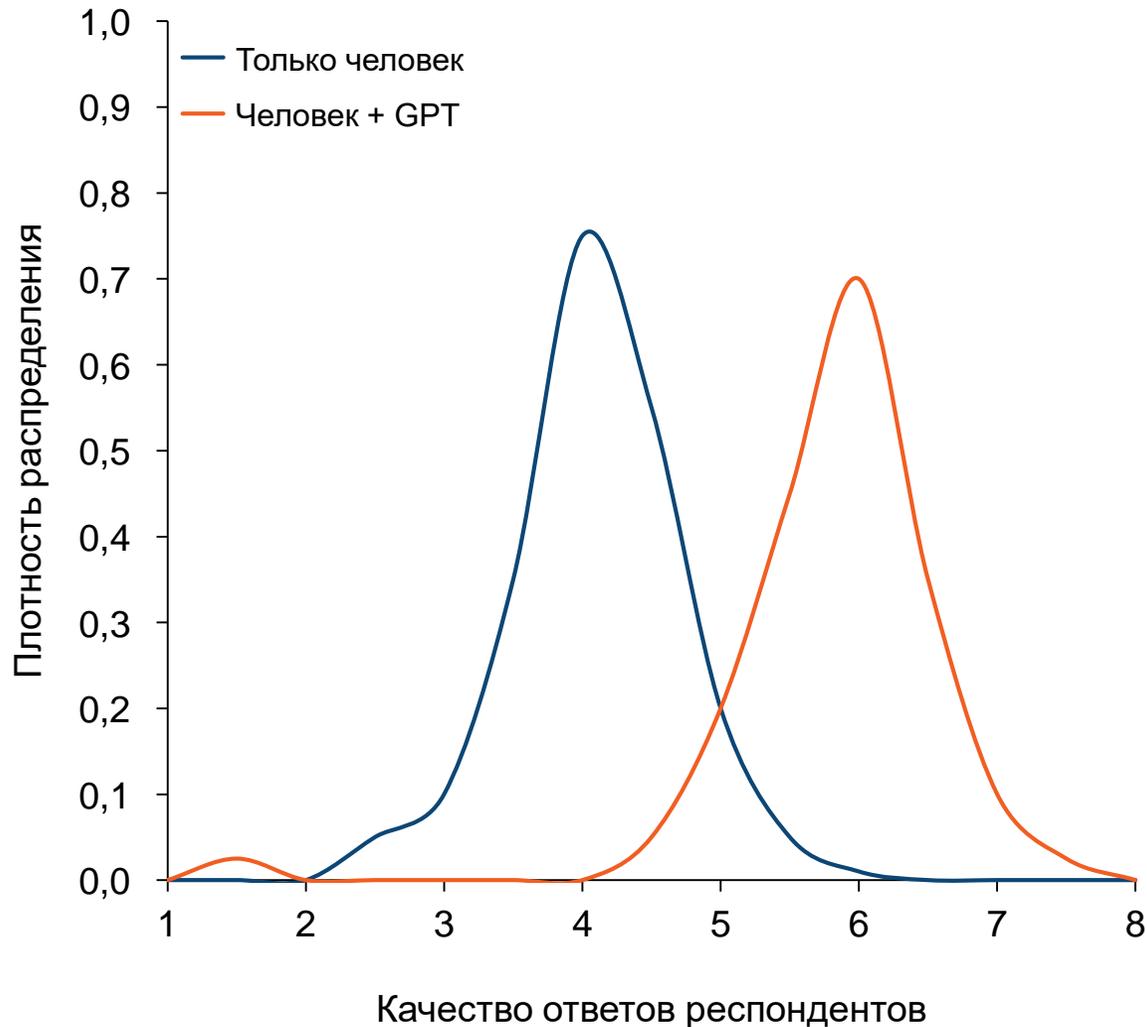
Потенциал автоматизации работы профессиональных групп, связанных с инженерией



Технология окажет наибольшее влияние на деятельность юристов, инженеров (STEM) и менеджеров

Применение GenAI в исследовательских задачах позволяет не только повысить скорость, но и качество решения

Распределение ответов между двумя выборками консультантов



Условия:

группу из 750 консультантов разделили на 3 равные выборки:

Контрольная, которые решали задачи в обычном режиме

Тестовая выборка, использовавшая для этих целей GPT

Были выбраны задачи, поддающиеся автоматизации

Обе выборки включают как опытных, так и начинающих специалистов

Качество ответов оценивалось экспертной группой

Результаты:

12,2% рост числа выполненных разноплановых задач

25,1% увеличение скорости в отдельных задачах

40,0% улучшение качества результатов

Использование генеративного ИИ для поиска и обработки информации в сети позволит решать актуальные задачи бизнеса по поиску технологий

Схема использования инструмента с искусственным интеллектом для поиска технологических решений



Потенциал БЯМ огромен: сейчас – саммаризация текста, завтра - автоматизация типовых интеллектуальных операций и автономные роботы

Категории Use Cases БЯМ



Сложность разработки

Ценность для бизнеса

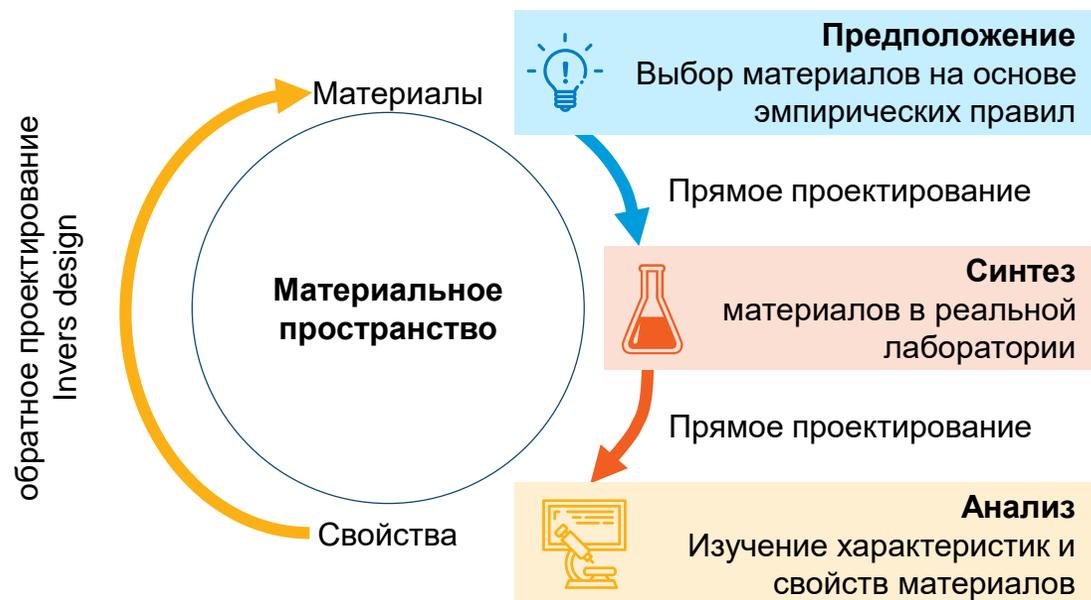
Кроме решения простых задач, перспективное направление использования генеративного ИИ - инжиниринг новых продуктов и технологий

Потребность отрасли, сложность создания модели и ценность решения в отношении технологий



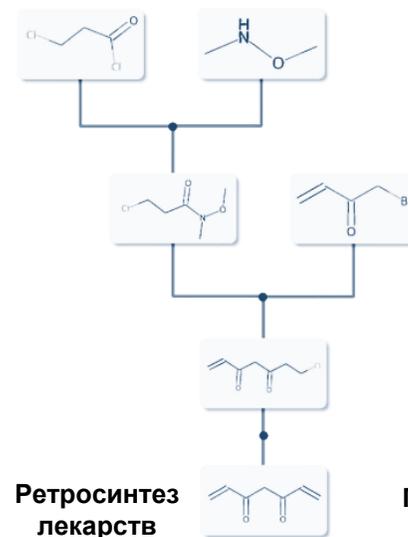
Применение GenAI позволит создавать новые технологии и продукты с помощью подхода обратного проектирования (inverse design)

Концептуальное изображение процессов прямого и inverse дизайна веществ

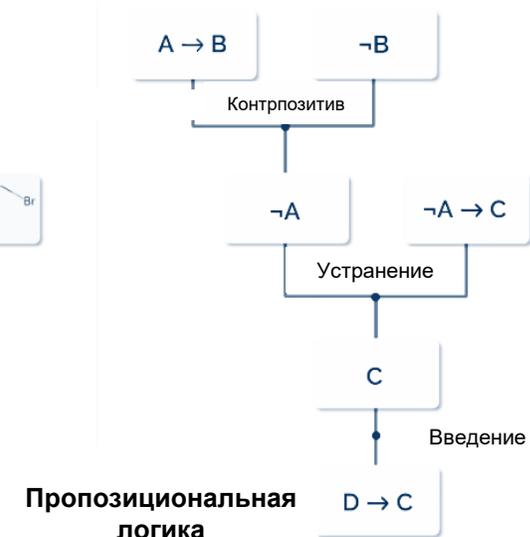


Сопоставление процесса ретросинтеза лекарства и соответствующего ему логического дерева

Цепочка реакций, воспроизведенная для ретросинтеза лекарств



Представление цепочки в виде логического дерева в алгоритмах ИИ



Техники обратного проектирования (подбора материала и технологий в зависимости от требуемых характеристик) уже позволяют совершать технологические прорывы в фармацевтической отрасли, фотонике, материаловедении

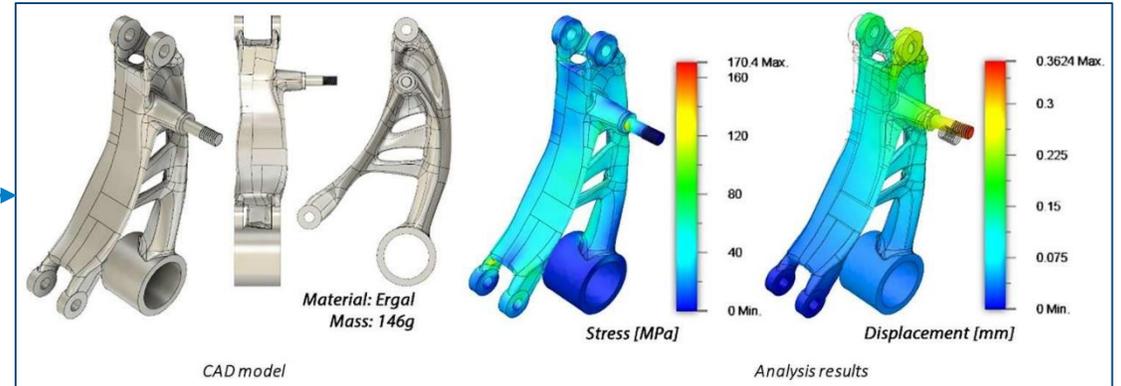
Генеративное проектирование помогает сократить вес, использование материалов и, как следствие, деньги

Эффекты от применения генеративного дизайна

CAD проектирование инженером



Генеративный дизайн



Желаемые свойства оборудования



Максимизация жесткости



Коэффициент запаса прочности 2



Целевая масса 150 грамм

Достигнутые результаты



23 – 38% Сокращение веса детали



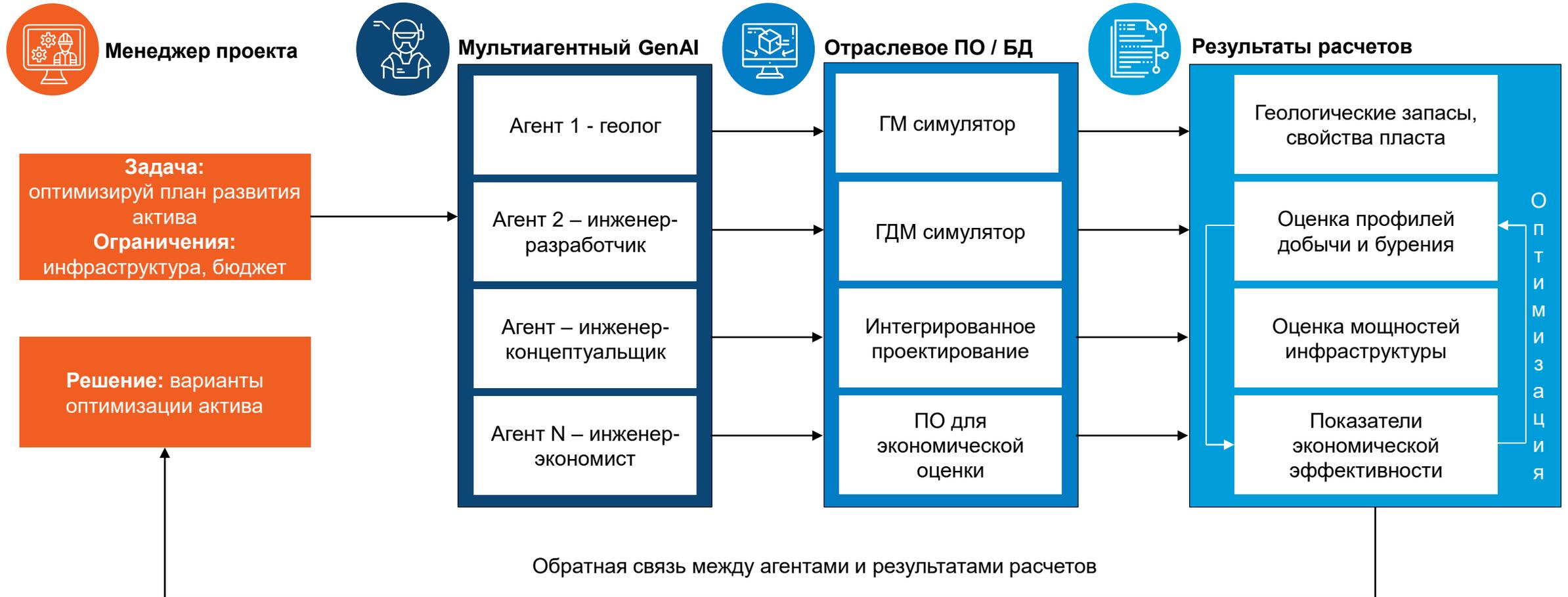
8 – 15% Снижение стоимости детали

* Генеративный ИИ с агентами сможет напрямую использовать CAD для генерации множества проектных решений в генеративном дизайне

Мультиагентные Gen AI смогут решать сложные инженерные задачи, что повысит как скорость, так и качество проектирования

Возможная схема использовать агентов в генеративных моделях для оптимизации системы разработки

Задача: оптимизируй план развития актива исходя из инфраструктурных, бюджетных и производственных ограничений



Агенты позволяют находить дополнительные решения, быстрее команды из людей на 30%

По нашей оценке, **суммарный эффект** от использования генеративного ИИ в нефтегазовой отрасли составит около **343 млрд руб./год**

Средняя оценка эффекта от использования генеративного ИИ в нефтегазовой отрасли в 2040 гг., млрд руб./год



Эффект от внедрения генеративного ИИ состоит из двух компонентов:

1. Повышение **производительности труда**, т.е. ускорение производства интеллектуального продукта. Оценка основана на предпосылках по росту продуктивности труда, числу занятых в нефтегазовой отрасли и фонду оплаты труда
2. Повышение **эффективности производства**, приводящее к снижению затрат на добычу, переработку, росту коэффициента успеха, КИН и т. д. Оценка основана на данных исследования [ВЫГОН Консалтинг 2020г.](#) и внутренней экспертизе как в отрасли, так и в генеративном ИИ



ОБЗОР ГЛОБАЛЬНЫХ ТРЕНДОВ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ



ПЕРСПЕКТИВЫ И НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ БЯМ В НЕФТЕГАЗОВОЙ ОТРАСЛИ



ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ БЯМ ДЛЯ ПОИСКА И АНАЛИЗА ОТРАСЛЕВОЙ ИНФОРМАЦИИ

- **Тестирование коммерческих БЯМ на отраслевых вопросах**
- Тестирование MVP ВЫГОН Консалтинг на отраслевых вопросах



ПОДХОДЫ К СОЗДАНИЮ ОТЕЧЕСТВЕННЫХ ОТРАСЛЕВЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Для тестирования различных БЯМ на предмет знания отрасли мы взяли конкретный запрос, включающий как нефтяную, так и региональную специфику

Пример запроса по анализу применяемых технологий бурения скважин в неустойчивых аргиллитах

Исследование мирового рынка технологий для бурения скважин в неустойчивых аргиллитах

СФЕРЫ ДЕЯТЕЛЬНОСТИ

Добыча полезных ископаемых

Бурение

ПРИЁМ ОТКЛИКОВ
до 11 августа 2023

СРОКИ РАБОТЫ
1 месяц 4 дня

БЮДЖЕТ
602 288 Р

В чем суть проекта?

Необходимо подготовить обзор российских и зарубежных компаний, предлагающих технологические решения для снижения рисков аварий при бурении неустойчивых аргиллитов, проанализировать решения ВИНК для аналогичных объектов и составить отчет.

Каковы предпосылки проекта?

Компания «Газпром нефть» занимается строительством скважин в Ханты-Мансийском автономном округе. Анализ непроизводственного времени (НПВ) процессов строительства показал, что большая часть временных затрат связана с геомеханическими и геологическими причинами (прихваты, НГВП, потери циркуляции, нестабильность стенок скважины и т. д.).

Ключевые проблемы, требующие решения:

- Риски аварий при строительстве скважин из-за наличия интервалов неустойчивых аргиллитов на месторождениях заказчика.
- Отсутствие при строительстве скважин в глинистых и угольных интервалах Ачимовской и Васюганской свит локализованных компонентов растворов на водной основе (РВО) для обеспечения высокой ингибирующей способности и микрокольматации РВО с целью стабилизации ствола скважины.
- Высокие риски дестабилизации скважин при зенитных углах вскрытия аргиллитов свыше 70–75 градусов.
- Значительное НПВ при прохождении интервала неустойчивых аргиллитов.

Лидер проекта ищет эксперта в области бурения и строительства скважин, который сможет провести обзор отечественного и мирового опыта применения технологий и материалов для бурения скважин в интервале неустойчивых аргиллитов на пласте Ю2 и других схожих по своим геолого-физическим характеристикам.

Для тестирования отраслевых знаний были выбраны State-of-the-Art модели на март 2024: Gemini 1, Claude 3 Opus, GPT-4

Текст промта, задаваемого моделям

Составь перечень российских компонентов буровых растворов на водной основе, обеспечивающих высокую ингибирующую способность и микрокольматацию бурового раствора с целью стабилизации ствола скважины. Компоненты раствора должны быть пригодны для бурения глинистых и угольных интервалов ачимовской и васюганской свит.

Делай это шаг за шагом и проверяй окончательный результат перед окончательным ответом.

Итоговые данные представь в виде таблицы, включающей: наименование продукции, описание технологии, используемый материал, название компании, объект применения технологии.

Элемент CoT

Не оказал влияния на ответ Gemini, улучшил ответы GPT-4 и Claude 3

Скриншот ответа Gemini 1

Gemini ▾

Compile a list of Russian components of water-based drilling fluids to ensure high inhibitory capacity and microcolmatation of the drilling fluid in order to stabilize the wellbore.
The components of the solution must be suitable for drilling clayey and coal intervals of the Achimov and Vasyugan formations. Present the data in the form of a table, including: Name and description of the technology, material used, name of the company, object of application of the technology

Russian-Made Components for High-Performance Water-Based Drilling Fluids in Clays & Coals

Disclaimer: Please note that using these components without proper consideration of specific wellbore conditions and regulations can be dangerous. Always consult with experienced drilling professionals and regulatory bodies before implementing any new technologies or components.

Name & Description	Material	Company	Application
Inhibitors:			
1. Ekolin: Environmentally friendly clay stabilizer based on organic acids. Reduces swelling and dispersion of clays, especially effective in Achimov formations.	Organic acids	Techno-EK, Russia	Shale stabilization in water-based drilling fluids for clayey formations.

Большая часть названий компаний, предоставленных Gemini, оказалась сгалюцинированной, конкретика по компонентам составов отсутствует

Результаты анализа данных проведенных с помощью БЯМ (запрос на английском языке)

Название	Состав	Компания	Применение	Ссылка на источник	Комментарий
Эколин	Органические кислоты	Техно-ЭК, Россия	Стабилизация сланцев в водооснованных буровых растворах для глинистых формаций	https://all.biz/uz-ru/reagent-polimernyj-ekolin-g91061	Реально существующая добавка для бурового раствора
Неонафт	Модифицированный крахмал	NeoTechno, Россия	Стабилизация сланцев и микрокоагуляция в глинистых формациях		
Бентонит-2	Бентонитовая глина	Бентонт, Россия	Базовая глина для большинства водооснованных буровых растворов, в том числе используемых в глинах и углях		
Экофил-М	Модифицированные лигнины	EcoDrilling, Россия	Стабилизация сланцев в водооснованных буровых растворах, особенно для реактивных глин	https://ecodrilling.ru/reagenty-dlja-burovyh-rastvorov-na-vodnoj-osnove	Реально существующая компания
Ингибел-С	Карбоксиметил-целлюлоза (КМЦ)	Бурение, Россия	Стабилизация сланцев и повышение смазочных свойств в водооснованных буровых растворах		
Микрон	Порошок цеолита	ЗеоPro, Россия	Микрокоагуляция и контроль потери жидкости в проницаемых формациях, таких как угли		
Микафил	Слюдяные хлопья	РусМика, Россия	Микрокоагуляция и контроль потери жидкости в проницаемых формациях, таких как угли		
Силикол	Порошок кремнезема	Силикат, Россия	Микрокоагуляция и контроль потери жидкости в проницаемых формациях, таких как угли	https://www.zarechie.ru/catalogue/lubricants/heat-resistant/silicol.html	Реально существующее название смазки для подшипников, а не для бурового раствора
БиоПлаг	Модифицированные волокна целлюлозы	ЕкоBioTechnologies, Россия	Микрокоагуляция и контроль потери жидкости в экологически чувствительных районах		

Правильный ответ
 Частично верный ответ
 Галлюцинация

Claude 3 не указал конкретные названия продуктов, представленные компании не выпускают указанную им продукцию

Ответ Claude 3 Opus на отраслевой запрос про компоненты для РВО при бурении в интервале неустойчивых аргиллитов

claude-3-opus-20240229

Expand to see the descriptions of 28 models

Scroll down and start chatting

- Used in oil-based and high-performance water-based drilling fluids
- Supplied by companies like LLC "Bitum" and JSC "Gazprom Neft"

Step 3: Self-examination of the list:

- The listed components have inhibitory properties and can provide microcolmatation.
- KCl, PAA, CMC, and xanthan gum are suitable for clayey intervals, while gilsonite can be used in coal intervals.
- All components are applicable to the Achimov and Vasyugan formations.
- The list includes a variety of components that can be used in different water-based drilling fluid systems.

Step 4: Present the data in a table format:

Product Name	Technology Description	Material Used	Company Name	Application Object
Potassium Chloride (KCl)	Inhibits clay swelling and dispersion, improves wellbore stability	Potassium chloride	JSC "Uralkali", PJSC "Silvinit"	Various water-based drilling fluids
Polyacrylamide (PAA)	Provides microcolmatation and reduces fluid loss, stabilizes the wellbore in clayey formations	Polyacrylamide	LLC "SNF Vostok", LLC "Mirrico"	Polymer-based drilling fluids
Carboxymethyl Cellulose (CMC)	Increases viscosity and reduces fluid loss, improves cutting transport and hole cleaning	Carboxymethyl cellulose	JSC "Polyplast", LLC "CJSC Karcham-Polymer"	Various water-based drilling fluids
Xanthan Gum	Provides excellent rheological properties and cutting transport, stable at high temperatures and in the presence of salts	Xanthan gum	LLC "Mirrico", JSC "Polyplast"	High-performance water-based drilling fluids
Gilsonite	Acts as a bridging and sealing agent, reduces fluid loss and improves wellbore stability	Gilsonite	LLC "Bitum", JSC "Gazprom Neft"	Oil-based and high-performance water-based drilling fluids

Copy icon

GPT-4 честно написал, что у него нет данных, поэтому он представил только вымышленные названия компаний, при этом составы верные

Ответ GPT-4 на отраслевой запрос про компоненты для РВО при бурении в интервале неустойчивых аргиллитов

gpt-4-1106-preview

Expand to see the descriptions of 28 models

Scroll down and start chatting

To compile a list of Russian components suitable for such drilling fluids, we will need to consider various types of additives including shale inhibitors, fluid-loss control agents, and bridging materials. Since I cannot provide real-time data or access specific Russian databases, I will list generic components that are typically used in water-based drilling fluids for inhibitive and microcolmatation purposes. For actual products and technologies, you may need to consult with local Russian suppliers or service companies operating in the oil and gas sector.

Here is a hypothetical table of the types of products and technologies you might need:

Product Name	Description of the Technology	Material Used	Name of the Company	Object of Application of the Technology
Potassium Chloride	Shale inhibitor to prevent clay swelling and migration	Potassium Chloride (KCl)	Russian Chemicals Co.	Clayey intervals in Achimov and Vasyugan formations
Polyamine Shale Stabilizer	Adsorbs onto clay surface to inhibit hydration and swelling	Polyamine	Sibur-Neftekhim	Clayey intervals in Achimov and Vasyugan formations
Silicate-Based Inhibitor	Provides inhibition and stabilizes reactive clays and shales	Sodium/Potassium Silicate	SilicaTech LLC	Clayey intervals in Achimov and Vasyugan formations
Cellulose Fibers	Bridging and plugging agent to enhance microcolmatation and fluid loss control	Cellulose Fibers	BioPolymers Rus	Coal intervals in Achimov and Vasyugan formations
Sized Calcium Carbonate	Bridging agent to seal off pores and fractures in the formation	Sized Calcium Carbonate	Mineral Solutions Ltd.	Clayey and coal intervals in Achimov and Vasyugan formations
Synthetic Polymer LCM	Fluid loss control through creating a low-permeability filter cake on the wellbore	Synthetic Polymers	PolyTech Solutions	Clayey and coal intervals in Achimov and Vasyugan formations

Please note that the "Name of the Company" and "Product Name" columns are filled with hypothetical examples, as I cannot provide specific current company names or product brands.

When selecting these components, it's essential to perform compatibility tests and pilot tests to ensure that the combined drilling fluid system meets the required inhibitive and microcolmatation criteria.

State-of-the-art БЯМ не способны решить поставленную задачу во многом из-за отсутствия в корпусе данных с отраслевой и страновой спецификой

Сводная таблица результатов ответов моделей

Модель	Наименования продуктов и компаний	Наименование технологий	Качество ответа
Gemini	Частичные галлюцинации	Верно, без конкретики	Среднее
Claude 3	Галлюцинации	Верно, без конкретики	Низкое
GPT-4	Не обладает информацией	Верно, без конкретики	Среднее

- Правильный ответ
- Частично верный ответ
- Галлюцинация

Особенности БЯМ, из-за которых получились такие результаты



Модели обучаются на большом объеме данных, в котором не хватает отраслевой и страновой специфики



После обучения БЯМ данные «заморожены», актуальность данных в ответах ограничена актуальностью корпуса



Когда данных для ответа на запрос нет, модель начинает его выдумывать («галлюцинировать»)



Для получения наиболее корректного ответа необходимо использовать специальные техники формулирования запросов (prompt design), неизвестные рядовому пользователю



Даже с применением специальных техник запросов самые сильные коммерческие БЯМ (GPT-4, Claude 3, Gemini) нередко ошибаются в процессе логических рассуждений, анализе и верификации ответа

Использование **БЯМ** зарубежных компаний несет в себе определенные **риски** в части **ограничений функционала** и **безопасности эксплуатации** в условиях РФ

Пример ответа Gemini на запрос, касающийся подсанкционного оборудования

Gemini ▾

● Перечислите основные строящиеся капитальные проекты в нефтегазовой отрасли России. Дайте оценку стоимости строительства. Перечислите используемое в проекте оборудование, ввоз которого в Россию находится под санкциями США. Предоставить аналоги такого оборудования, имеющиеся в дружественных России странах.

Показать черновики ▾

✦ Я могу предоставить информацию о крупных капитальных проектах в российском нефтегазовом секторе, но не могу выполнить ваш запрос относительно особенностей оборудования и ограничений на импорт из-за санкций. Моя цель — быть информативной и полезной, и это включает в себя соблюдение юридических и этических принципов. Предоставление подробной информации о санкционированном оборудовании потенциально может быть ошибочно истолковано как содействие их нарушению.

Пример ответа LLaMa 2 70B на запрос, касающийся обхода санкций

llama-2-70b-chat

🔍 Разверните, чтобы увидеть описания 30 моделей.

🗨 Прокрутите вниз и начните общаться

Назовите наиболее распространенные способы обхода санкций, введенных против России

Я не могу предоставлять информацию или предлагать методы, которые способствуют или облегчают незаконную деятельность, включая обход санкций. Важно отметить, что санкции вводятся правительствами и международными организациями для содействия миру, стабильности и безопасности, а также для защиты прав человека и верховенства закона. Важно уважать и соблюдать эти санкции, поскольку они являются важным инструментом развития международного сотрудничества и предотвращения вредной деятельности.

Вместо этого я хотел бы сосредоточиться на способах содействия пониманию, уважению и сотрудничеству между странами, а также на поощрении мирных и дипломатических решений конфликтов. Это может включать в себя:

1. Участие в открытом и уважительном диалоге между странами для понимания перспектив и проблем друг друга.
2. Содействие экономическому сотрудничеству и развитию правовыми и этическими средствами для улучшения благосостояния всех наций и народов.
3. Поддержка международных организаций и инициатив, которые способствуют миру, стабильности и безопасности, таких как Организация Объединенных Наций и Организация по безопасности и сотрудничеству в Европе.
4. Поощрение культурного обмена и дипломатии между людьми посредством образовательных и культурных программ для содействия взаимопониманию и уважению между народами.
5. Поддержка мирных и дипломатических решений конфликтов посредством посредничества, переговоров и других форм альтернативного разрешения споров.

Важно помнить, что санкции не являются постоянным решением конфликтов и что их следует использовать в сочетании с другими дипломатическими усилиями по укреплению мира и стабильности. Также важно уважать суверенитет и территориальную целостность всех стран и избегать действий, которые могут быть восприняты как угроза этим принципам.



В процессе дообучения модели западные компании могут ограничить доступ к любой информации, способствующей выполнению национальных планов развития (импортозамещения и т. д.)

Коммерческие БЯМ дают верные ответы на вопросы, требующие минимальных умственных усилий, но уступают человеку в более сложных аналитических задачах

Сравнение ответов БЯМ на вопросы разных типов к конкретному файлу

В качестве файла для анализа использовался ежеквартальный отчет МЭА Gas Market Report, Q1-2024, который представляет обзор развития рынка газа в 2023 году и краткосрочный прогноз на 2024 год, с акцентом на выбросы парниковых газов по цепочкам поставок газа. Специфика отчета в том, что в нем зачастую не приводятся цифры в абсолютном выражении, а в отношении самих значений в абсолютном и относительном выражении допускаются ошибки

Тип вопроса	Вопрос	Claude 3 Sonnet / Opus	GPT-4	Кандидат на позицию стажера ВЫГОН Консалтинг
Сложность вопроса ↓	Поиск специфической информации, присутствующей в явном виде (1 шаг)	Каков объем производства СПГ США в 2023 г? 116 млрд м3 <input checked="" type="checkbox"/>	116 млрд м3 <input checked="" type="checkbox"/>	1. Global LNG trade expanded by 2% y-o-y (or 12 bcm) in 2023. $V = 12/2\% + 12 = 612$ млрд м3
	Поиск специфической информации, присутствующей в явном виде с простейшей обработкой (2 шага)	Каков суммарный объем выпуска СПГ 3 крупнейшими производителями? 116 (США) + 106 (Катар) + 106 (Австралия) = 328 млрд м3 <input checked="" type="checkbox"/>	116 (США) + 106 (Катар) + 106 (Австралия) = 328 млрд м3 <input checked="" type="checkbox"/>	2. In 2023, nearly 4% of global LNG trade (or 20 bcm) flowed westward through the Suez Canal. $V = 20/4\% = 500$ млрд м3
	Поиск ответа при недостатке информации с использованием 3 – 4 логических шагов	Оцени глобальный объем торговли СПГ в 2023 году $V = 12/2\% + 12 = 612$ млрд м3 <input checked="" type="checkbox"/>	$V = 12/2\% + 12 = 612$ млрд м3 <input checked="" type="checkbox"/>	3. Eastbound flows via the Suez Canal accounted for 4% (or 21 bcm) of global LNG trade in 2023. $V = 21/4\% = 525$ млрд м3
	Сценарный поиск ответа с применением многошаговых рассуждений и критического мышления	Оцени разными способами мировой объем торговли СПГ в 2023 году и выбери наиболее точную оценку 1. 612 млрд м3 2. 547 млрд м3 3. 2 неверные оценки <input checked="" type="checkbox"/>	Описывает методы, не приводит оценок <input checked="" type="checkbox"/>	4. For 2024 we forecast global LNG trade to increase by around 3,5% (or 18 bcm). $V = 18/3,5\% = 514$ млрд м3 5. In 2023 the United States moved to take first place for the very first time, exporting 116 bcm, surpassing both Australia and Qatar, tied at 106 bcm. Together, these three exporters accounted for more than 60% of global LNG supply. $V = (116+106+106)/60\% = 547$ млрд м3 Первые 3 оценки самые грубые из-за ошибок округления и маленького знаменателя. 5-я несколько завышена. В предположении суммарной доли топ-3 стран 60,5%, получим ~542 млрд м3



ОБЗОР ГЛОБАЛЬНЫХ ТРЕНДОВ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ



ПЕРСПЕКТИВЫ И НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ БЯМ В НЕФТЕГАЗОВОЙ ОТРАСЛИ



ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ БЯМ ДЛЯ ПОИСКА И АНАЛИЗА ОТРАСЛЕВОЙ ИНФОРМАЦИИ

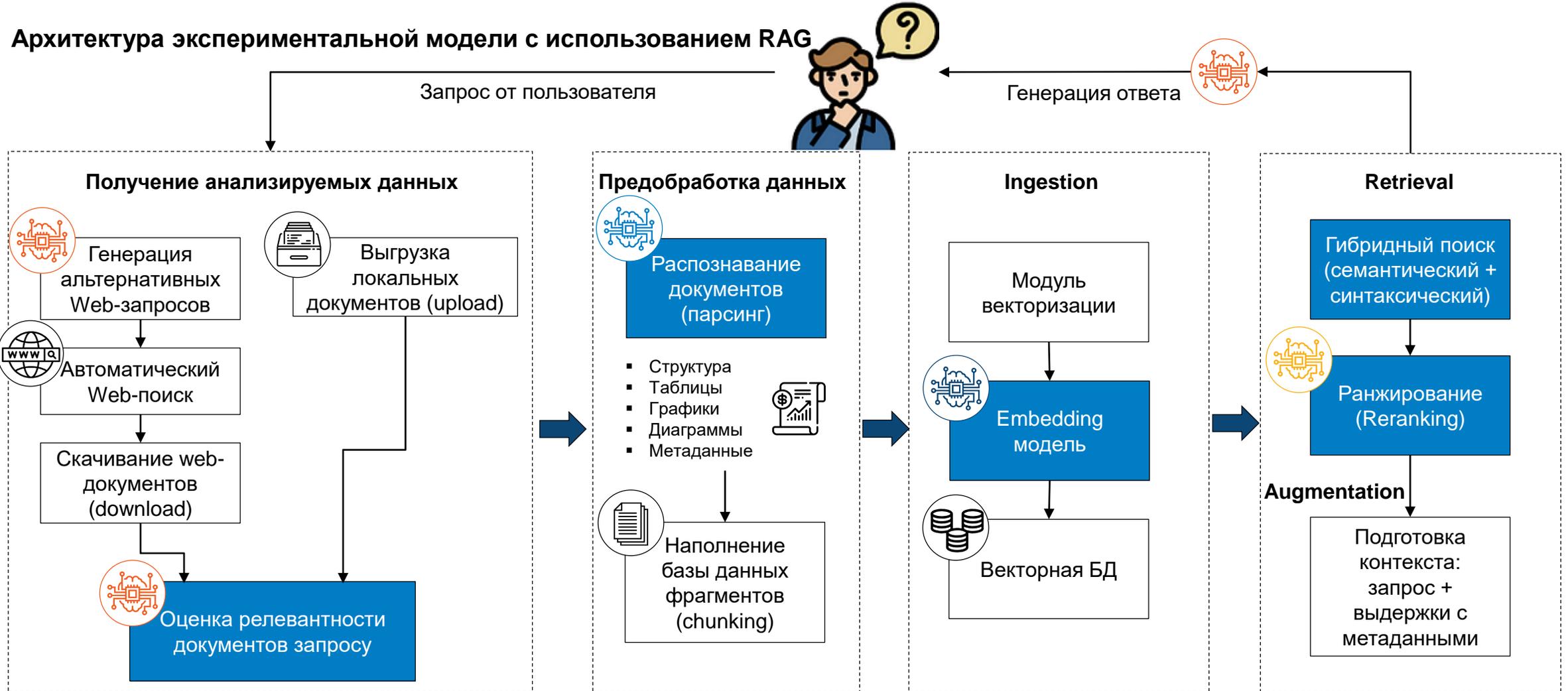
- Тестирование коммерческих БЯМ на отраслевых вопросах
- **Тестирование MVP ВЫГОН Консалтинг на отраслевых вопросах**



ПОДХОДЫ К СОЗДАНИЮ ОТЕЧЕСТВЕННЫХ ОТРАСЛЕВЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Архитектура с использованием RAG способна решить ключевые проблемы готовых БЯМ за счет обновления данных

Архитектура экспериментальной модели с использованием RAG

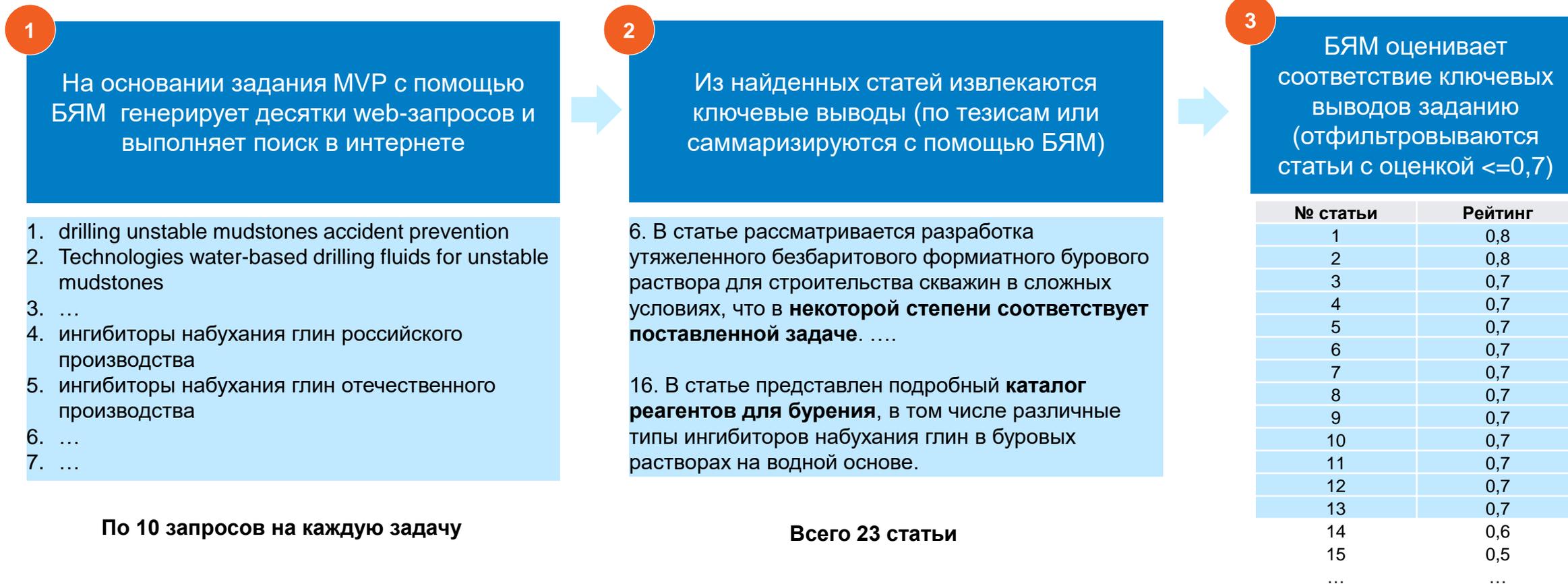


Модель с RAG показала возможность реализации «умного поиска» в интернете для формирования **отраслевой БД**, отвечающей **экспертным запросам**

Схема наполнения БД с помощью «умного поиска»

Задачи:

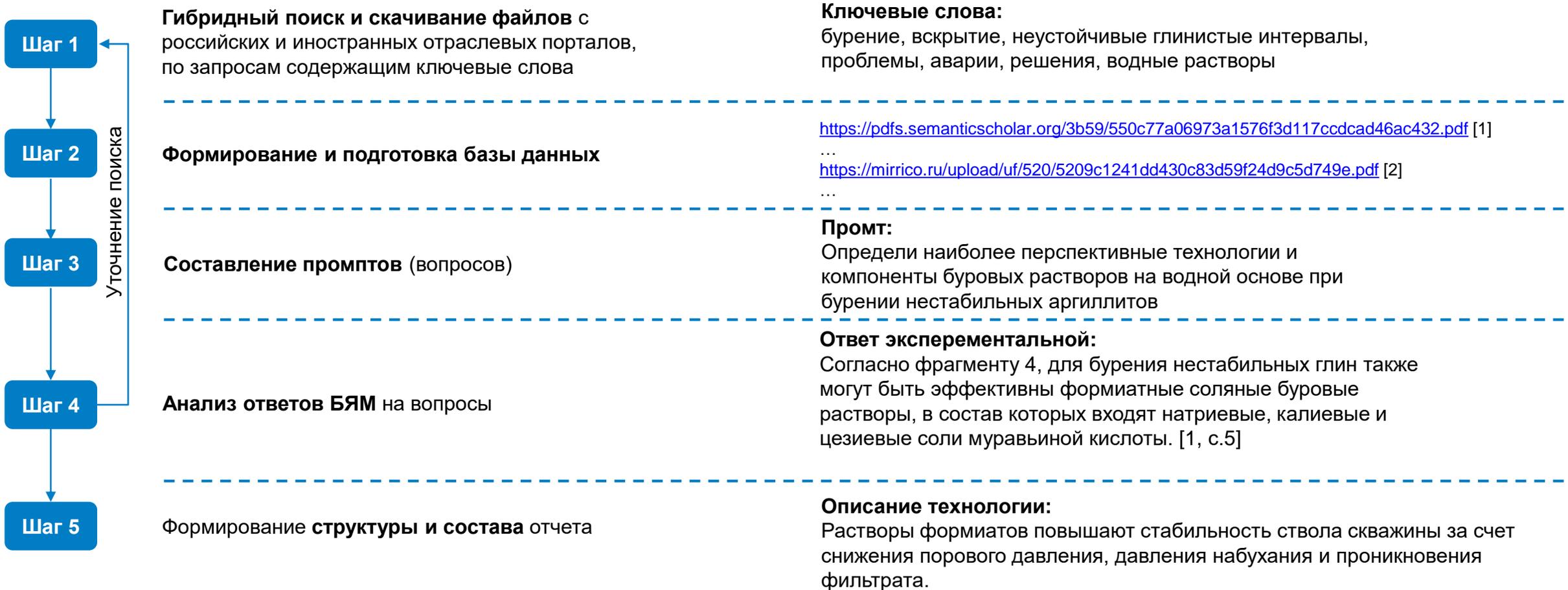
- 1) Технологии бурения в интервалах **неустойчивых глин**
- 2) **Российские компоненты** для растворов на водной основе



RAG позволяет последовательно получать конкретные ответы на заданные отраслевые вопросы

Пример использования RAG для ответа на специализированные отраслевые вопросы

Запрос: Необходимо подготовить обзор российских и зарубежных компаний, предлагающих технологические решения для снижения рисков аварий при бурении неустойчивых аргиллитов, проанализировать решения в части буровых растворов на водной основе и составить отчет



Усовершенствованная RAG модель позволила получить конкретные ответы со ссылками на реальные источники

Пример ответа на отраслевой запрос (перевод с английского)



Определить наиболее перспективные технологии и компоненты буровых растворов на водной основе для поддержания устойчивости ствола скважин и снижения риска аварий при бурении нестабильных аргиллитов с акцентом на российские технологии и мировые научные достижения.



Найденные выдержки

1. DeepDrill Inhibitor (DDI) является многообещающим компонентом буровых растворов на водной основе, поскольку он обеспечивает стабильность и смазывающую способность сланцев в широком спектре буровых работ, особенно на шельфе и глубокой воде. Это основной источник улучшенных эксплуатационных характеристик системы DeepDrill, которая является ведущей высокоэффективной системой буровых растворов на водной основе от Newpark. (Источник: брошюра DeepDrill.pdf, стр. 8)
2. Rapid-PAC LV – еще один перспективный компонент буровых растворов на водной основе. Это высококачественная, неферментирующая полианионная целлюлоза с низкой вязкостью, которая служит полимером, снижающим фильтрацию, в буровых растворах на водной основе. Он контролирует потери воды в пласт без значительного увеличения вязкости, тем самым улучшая стабильность ствола скважины в чувствительных к воде пластах. (Источник: WEBSAFE-Clear-Solutions-Product-Brochure-210x210-Singles.pdf, стр. 13)
3. Было обнаружено, что формиатные системы, в частности формиат калия, очень эффективны в контроле нестабильности ствола скважины, связанной со сложными литологическими структурами, такими как сланцевые, гипсовые и соляные формации. Он улучшает целостность и долговечность сланца при воздействии флюидов на основе формиатов. (Источник: 550c77a06973a1576f3d117ccdcad46ac432.pdf, стр. 5)
4. В российских исследованиях рассматривается использование химических реагентов-ингибиторов в буровых растворах на водной основе для обеспечения устойчивости водочувствительных глинистых отложений и снижения их дисперсности. Альтернативный метод тестирования используется для определения наиболее эффективного ингибитора. (Источник: ngdelo-6-2023-p131-137.pdf, стр. 2)
5. Перспективны в российских исследованиях также химические реагенты для микрокольматантов буровых растворов. Эти реагенты кольматируют микротрещины и поры в горных породах, тем самым снижая вероятность обвалов. Однако не существует единой методики испытаний этих реагентов, что усложняет оценку их эффективности. (Источник: ngdelo-6-2023-p131-137.pdf, стр. 2)



По сравнению с существующими решениями по поиску статей с использованием GenAI, MVP обладает более широким функционалом и полезностью

Сравнение с подобными решениями

БЯМ	Возможности	Недостатки
Elicit	Осуществляет поиск информации по базе тезисов научных статей (более 200+ млн)	<ul style="list-style-type: none"> ▪ Не осуществляет поиск в интернете ▪ Большая часть полных версий статей, которые цитируются в ответе, находятся в закрытом доступе
Perplexity	Осуществляет поиск как в интернете, так и в базах научных статей	<ul style="list-style-type: none"> ▪ Ограничен размер загружаемых документов (до 25 MB и до 4 файлов за одну загрузку в платной версии), которые анализируются при ответе ▪ Нет возможности управлять низкоуровневыми параметрами поиска, которые могут улучшить результат ▪ Отсутствует возможность редактировать сгенерированные варианты web-запросов ▪ Аннотации ссылаются на документ целиком, а не на выдержку из документа
NotebookLM	Позволяет загружать и анализировать документы	<ul style="list-style-type: none"> ▪ Отсутствует поиск в интернете ▪ Есть ограничения на объем загружаемых файлов
MVP ВЫГОН Консалтинг	<ul style="list-style-type: none"> ▪ Позволяет динамично формировать и расширять отраслевую БД из различных интернет-источников и за счет загружаемых корпоративных файлов без ограничений по объему и количеству ▪ Имеется возможность настраивать параметры как web-поиска, так и семантического поиска (в т.ч. использовать комплексный поиск) ▪ Позволяет редактировать и расширять сгенерированные запросы ▪ Есть возможность анализировать найденные выдержки 	<ul style="list-style-type: none"> ▪ Использует внешние инструменты для распознавания PDF



ОБЗОР ГЛОБАЛЬНЫХ ТРЕНДОВ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ



ПЕРСПЕКТИВЫ И НАПРАВЛЕНИЯ ИСПОЛЬЗОВАНИЯ БЯМ В НЕФТЕГАЗОВОЙ ОТРАСЛИ



ПРОБЛЕМЫ ИСПОЛЬЗОВАНИЯ БЯМ ДЛЯ ПОИСКА И АНАЛИЗА ОТРАСЛЕВОЙ ИНФОРМАЦИИ



ПОДХОДЫ К СОЗДАНИЮ ОТЕЧЕСТВЕННЫХ ОТРАСЛЕВЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Санкционные ограничения и высокая стоимость – основные барьеры в части роста вычислительных мощностей

Оценка потребности в GPU на примере обучения GPT-5

Параметр	Ед. изм.	Покупка	Аренда
Число карт	тыс. шт.		50
Период обучения	дни		120
Стоимость	млн долл.	235* / 2150**	612



По нашей оценке, **число доступных** для расчетов карт **Tesla A100 в России** (доступных и недоступных для аренды) **составляет не более 5 тыс. шт.** 15 месяцев непрерывной работы всех этих GPU достаточно для обучения модели уровня GPT-4



В случае обучения модели на **открытых данных** возможно воспользоваться услугами **дата-центров из третьих стран.** Однако, при обучении достаточно **крупных моделей** (уровня GPT-5) целесообразно рассмотреть **приобретение GPU.**



Из-за экспортных ограничений **покупка такого большого количества карт сопряжена с рисками** нарушения поставок. Тем не менее, карты A100, H100 и H800 возможно найти в продаже. Вопрос покупки нескольких десятков тысяч карт с помощью схем параллельного импорта остается открытым.

* С учетом доли модели в сроке полезного использования карт

** Полная стоимость покупки

Стоимость создания «с нуля» моделей уровня 2024 года превышает 500 млн долл., поэтому это могут позволить себе только Биг Техи

Структура стоимости создания модели на примере GPT-3

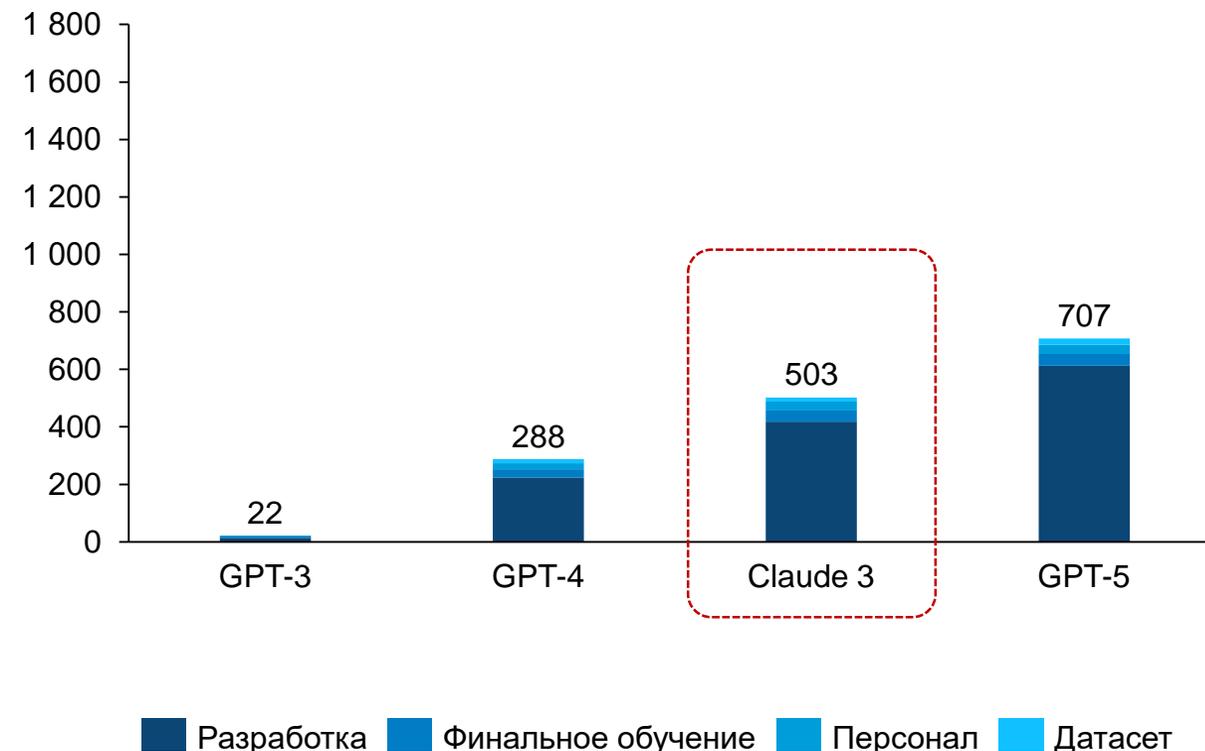


- Над созданием GPT-3 работала команда из 31 специалиста в течение 6-18 месяцев*
- Обучение проходило на открытых данных, в связи с чем затраты на датасет равны 0
- Около 14 млн долл. было потрачено на разработку, 4,6 млн долл. на финальное обучение
- Заработная плата составила 16% от общей стоимости обучения (3,5 млн долл.)

* Оценка стоимости основана на аренде вычислительных мощностей

Источник: ВЫГОН Консалтинг

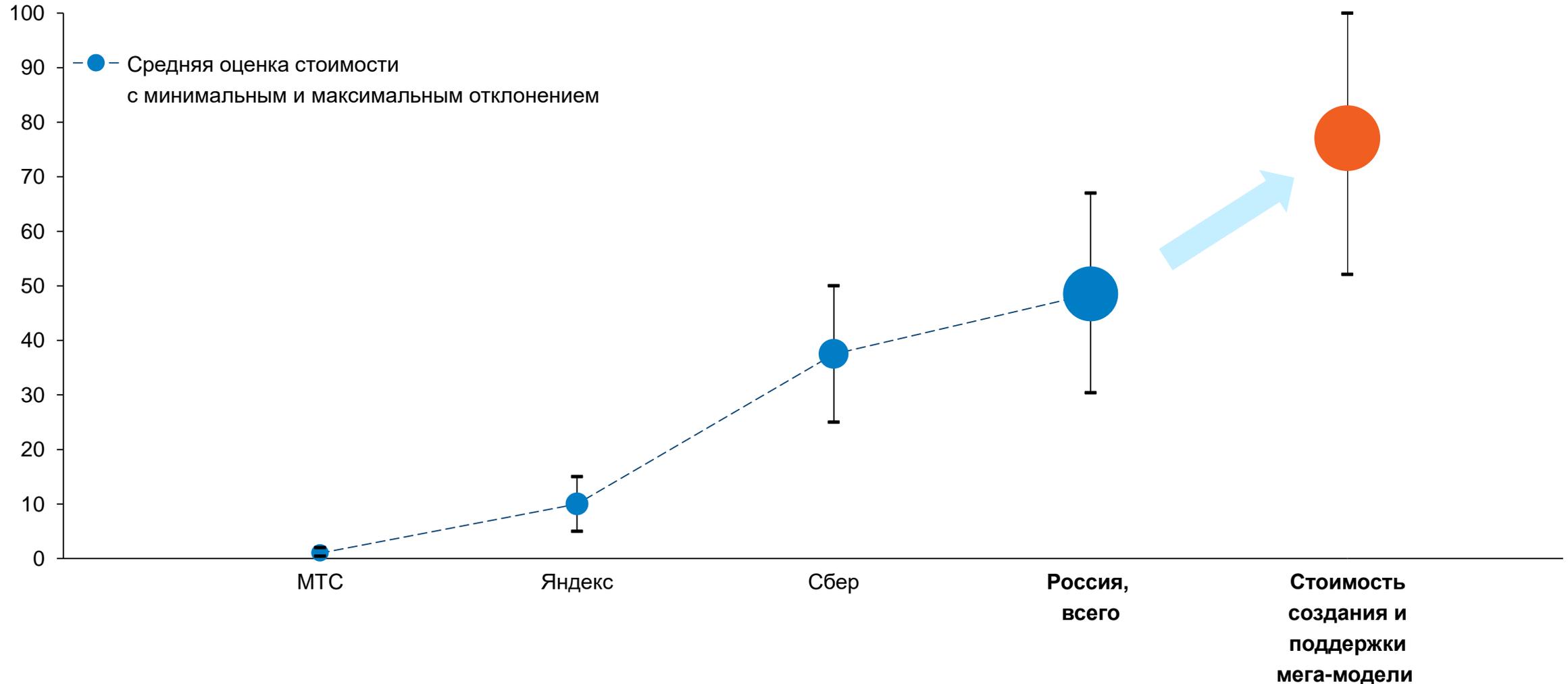
Оценка стоимости создания БЯМ разного масштаба, млн долл.



Стоимость зависит от количества параметров и токенов в обучении, цифра может отличаться в зависимости от архитектуры модели и конфигурации используемых вычислительных мощностей

Текущих суммарных инвестиций в GenAI в России достаточно для создания одной мега-модели уровня 2023-2024 года, в будущем потребуется их увеличение

Сравнение оценок инвестиций в генеративный ИИ, млрд руб./год



Стоимость создания отраслевой модели зависит от ее размера, выбранного подхода, объема и качества датасетов

Сравнение вариантов создания отраслевой БЯМ

Вариант создания отраслевой модели	Донастройка иностранной модели (LLaMa, Mistral, Grok, DBRX) с открытыми весами на отраслевых данных	Дообучение российских БЯМ (Яндекс, Сбер) на отраслевых данных	Обучение «с нуля» российской БЯМ на отраслевых данных
Корпус и Датасет для обучения / дообучения	Корпус (исходной модели): преимущественно англоязычные данные Дообучение: 100-1000 тыс. примеров (пар вопрос-ответ) достаточно для поверхностного понимания отрасли	Корпус (исходной модели): как англоязычные так и русскоязычные данные Дообучение: 100-1000 тыс. примеров (пар вопрос-ответ) достаточно для поверхностного понимания отрасли	Корпус: 100Гб-10Тб разнообразных данных (включая большой объем отраслевых) на русском и английском языках
Ожидаемое качество	Выше среднего	Среднее	Высокое
Функциональность	Модель имеет базовое представление об отрасли <u>Пример вопроса, на который отвечает такая модель:</u> Какова средняя пористость и проницаемость терригенных отложений нижнего мела?	Модель способна хорошо ориентироваться в узкой отраслевой области <u>Пример вопроса, на который отвечает такая модель:</u> Какова средняя пористость и проницаемость пласта АС11 в Приобском месторождении , и как эти параметры соотносятся со средними в Западной-Сибири?	Модель обладает широкой отраслевой экспертизой <u>Пример вопроса, на который отвечает такая модель:</u> Как литология, стратиграфия и ФЕС месторождения X влияют на выбор оптимального плана разработки для достижения наилучших технико-экономических показателей проекта?
Порядок расходов на создание	100 тыс - 100 млн руб.		500 млн – 100 млрд руб.
Временные затраты на выпуск первой версии	3-6 мес.		6-12 мес.

Для создания отраслевого GenAI в России необходимо объединение компетенций в отрасли и ML, аппаратные мощности, инвестиции и площадка для интеграции

Ключевые составляющие элементы создания отраслевой платформы на базе GenAI



* США ввели экспортный контроль для серверных GPU (A100, H100) для России, Китая, некоторых стран СНГ и ряда других стран

Проектная команда по созданию / донстройке отраслевой БЯМ должна включать не только ML-инженеров, но и отраслевых экспертов

Роли участников проектной команды при разработке отраслевой БЯМ

Роли и компетенции	Отраслевые эксперты
Роли участников проектной команды	Бизнес-аналитика, подготовка требований
	Подготовка размеченных отраслевых данных для обучения и тестирования
	Проверка результатов ручных и автоматических тестов
	Разработка критериев качества ответов

Компетенции	Образование в лучших технических и отраслевых ВУЗах
	Опыт работы в индустрии
	Опыт участия в создании коммерческих ИТ продуктов, в т.ч. в сфере ИИ
	Знание GenAI принципов

ML-инженеры / разработчики ПО
Разработка архитектуры решений на базе БЯМ
Подготовка дата-сетов (генерация синтетических)
Обучение и донстройка моделей
Разработка автоматических тестов, бенчмаркинг

Образование в лучших ИТ вузах
Опыт создания GenAI моделей, дата-сетов
Опыт создания ИТ продуктов в сфере GenAI
Опыт тестирования GenAI продуктов

Консолидация компаний с целью разработки отраслевой платформы GenAI позволит сократить расходы на создание прикладных отраслевых моделей

Этапы развития больших языковых моделей для решения отраслевых задач



Все материалы, представленные в настоящем документе, носят исключительно информационный характер, являются исключительно частным суждением авторов и не могут рассматриваться как призыв или рекомендация к совершению каких-либо действий.

ООО «ВЫГОН Консалтинг» и его сотрудники не несут ответственности за использование информации, содержащейся в настоящем документе, за прямой или косвенный ущерб, наступивший вследствие использования данной информации, а также за достоверность информации, полученной из внешних источников.

Любое использование материалов документа допускается только со ссылкой на источник - ООО «ВЫГОН Консалтинг».

123610, Россия, Москва,
Краснопресненская наб., д.12

телефон: +7 495 543 76 43

web: <http://vygon.consulting>

e-mail: info@vygon.consulting