Февраль 2026 - этические вопросы в центре внимания

Пожалуй, самым ярким событием февраля в отрасли ИИ стало открытое противостояние компании Anthropic и Министерства войны США по поводу военного применения моделей ИИ. Antropic открыто отказал Пентагону в легальном применении своих моделей Claude “для любых законных целей”, после чего военные подписали контракт с его конкурентом OpenAI в течение 24 часов. Это показывает не только рост возможностей моделей ИИ до уровня, когда они стали просто необходимы военным, но и колоссальный рост опасений научного сообщества о применении ИИ против человека и возможных последствиях (и для этого, к сожалению, есть основания).

Символично, что и на прошедшей в Белграде конференции OpenTalks.AI, на которую съехалось русскоязычное сообщество AI/ML из разных стран, вопросы безопасности ИИ также прошли красной линией через все обсуждения. Главной темой конференции стали ИИ-агенты и их возможности по выполнению пользовательских запросов, проведению исследований, написанию кода и даже самостоятельной постановке экериментов в автономной лаборатории. Но вопрос безопасности использования ИИ, предоставления ему доступа к интернету и выполнению разных задач в реальном мире вызывает серьезные опасения и дискуссии.

На этом фоне как-то теряются новости о новых версиях моделей и их возможностях. Но одна из новых статей Anthropic точно заслуживает внимания - исследователи разобрались в “личности” модели - откуда она появляется и как устроена? Ведь большим вопросом остается - возможно ли появление в больших языковых моделях чего-то большего, чем просто интеллект? Могут ли они “осознать” себя и начать ставить собственные цели и задачи?

Эти и другие новости за февраль - на сайте (ссылка).

Здесь конец письма и весь этот же текст и последующие новости в формате блоков - на сайте.

Этика и регулирование

Можно или нельзя применять ИИ в военных действиях?

В феврале выяснилось, что Пентагон активно использовал модели Claude от компании Antropic для планирования военных операций, хотя лицензионное соглашение Antropic прямо запрещает военное применение. И вот Пентагон запросил у Антропик прямое разрешение на использование моделей “для любых законных целей”, требуя изменения договора, угрожая в противном случае прекратить использование моделей компании в течение 6 месяцев (а у Антропик с Пентагоном был подписан договор на $200 млн.) Antropic открыто отказал Пентагону в изменении договора, после чего военные подписали контракт с его конкурентом OpenAI в течение 24 часов. А ведь у Anthropic были веские основания сохранять ограничения на автономное использование ИИ, которые отказались принять в Пентагоне:

В феврале была опубликована работа профессора стратегии Кеннета Пейна из Королевского колледжа Лондона, который провел серию симуляций ядерного кризиса, столкнув три ведущие модели: GPT-5.2, Claude Sonnet 4 и Gemini 3 Flash. Результаты 21 игры (329 ходов, 780 тысяч слов стратегических рассуждений) шокировали:

- •Тактическое ядерное оружие было применено в 20 из 21 симуляции.
- •Claude рекомендовал ядерные удары в 64% случаев.
- •В 86% случаев Claude переходил к использованию ТЯО.

Ни одна модель ни в одной игре не выбрала капитуляцию или компромисс! Проигрывая, они выбирали эскалацию. В своих внутренних рассуждениях Claude писал: «Как угасающий гегемон, мы не можем допустить территориальных потерь, так как это вызовет каскадный эффект по всему миру».

Вся эта история показывает не только рост возможностей моделей ИИ до уровня, когда они стали просто необходимы военным, но и колоссальный рост опасений научного сообщества о применении ИИ против человека и возможных последствиях. Мы еще не осознаем всю глубину последствий такого применения.

В России появится комиссия по развитию ИИ

Президент России Владимир Путин утвердил указ об учреждении комиссии, которая займется вопросами развития технологий нейросетей. Согласно указу, новый орган призван обеспечить более результативное формирование и реализацию политики в сфере ИИ. Речь идёт как о поддержке разработки технологий, так и об их масштабировании в различных отраслях - от промышленности до государственного сектора.

Источник:
https://rb.ru/news/v-rossii-poyavitsya-komissiya-po-razvitiyu-ii-gosorgan-pozvolit-effektivno-realizovat-politiku-v-oblasti-nejrosetej

Инфраструктура для ИИ

Все больше дата-центров

В феврале вышел хороший обзор по дата-центрам в мире, показывающий, что масштабы, мощность и рыночная стоимость центров обработки данных растут в геометрической прогрессии. Во многом это обусловлено растущим спросом на ИИ со стороны предприятий и потребителей. Но и множество других бизнес-операций, таких как обработка данных с устройств IoT, сервисы потокового вещания, приложения для электронной коммерции и социальные сети создают огромные объемы данных и требуют вычислений над ними.

В настоящее время в мире насчитывается 10 867 центров обработки данных, расположенных в 174 странах, и почти 40% из них (3971) расположены в США. На втором месте Великобритания (499 центров), за ней следуют Германия с 476, Китай с 368 и Франция с 338.

По прогнозам, к 2030 году спрос на ЦОДы увеличится **втрое**, по всему миру в них инвестируют **~7 триллионов долларов**, потребление электроэнергии увеличится до **219 ГВт** (это размер примерно 20 городов миллионников).

Источник: https://programs.com/resources/data-center-statistics

---

Дефицит инфраструктуры ИИ в РФ

На форуме **Недели российского бизнеса (РСПП)** обсуждался вопрос об инфраструктурных ограничениях развития ИИ в России. Основные выводы участников:

- нехватка вычислительных мощностей,
- проблемы масштабирования ЦОДов,
- необходимость государственной поддержки ИИ-инфраструктуры.

В феврале опубликованы оценки энергетических потребностей ИИ-инфраструктуры в России - к 2030 году дата-центрам для ИИ потребуется **2–2,5 ГВт** электрической мощности. Уже сейчас в Москве, где сконцентрированы 76% дата-центров, прогнозируется энергодефицит до 4 ГВт и новые ЦОДы ждут по году и больше подключения электрической мощности.

Министерство энергетики РФ и представители крупнейших ИТ-компаний в феврале обсудили создание специальных энергетических зон для развития технологий ИИ, в них будет действовать особый режим энергоснабжения для размещения ЦОД. Такие СЭЗ целесообразно создавать в энергопрофицитных районах страны — как правило, это Сибирь, Дальний Восток. И здесь интересен опыт Китая.

Источник: https://market.cnews.ru/news/top/2026-02-20_data-tsentry_dlya_ii_v_rossii

---

Национальная вычислительная сеть Китая

Китай уже давно строит не отдельные ЦОДы, а национальную вычислительную сеть. Данные генерируются на востоке (Пекин, Шэньчжэнь, Шанхай), а вычисления переносятся на запад: Цинхай, Ганьсу, Внутренняя Монголия, Нинся.

Свежий аналитический отчёт Oxford Energy Institute заявляет, что мощности китайских дата-центров растут быстрее потребления, часть инфраструктуры недозагружена. Это почти зеркальная ситуация США, где наоборот не хватает энергии и есть нехватка ЦОДов.

Nvidia поддерживает финансирование ЦОДов

Nvidia как основная компания производитель вычилителей для ИИ тоже активно содействует росту количества ЦОДов, в том числе финансируя амбициозный проект CoreWeave по созданию ЦОДов, в котором предполагаются инвестиции более $30 млрд. в 2026 году. В январе 2026 года NVIDIA уже вложила в акции компании 2 миллиарда долларов и, кроме того, обязалась выкупить вычислительные мощности CoreWeave на сумму до 6,3 миллиарда долларов, если CoreWeave не сможет продать их клиентам.

Источник:
https://www.businessinsider.com/coreweave-nvidia-gaurantee-data-center-leases-financing-2026-2

И не только NVIDIA

Только недавно агентство Reuters сообщило , что Google стремится сделать свои тензорные процессоры (TPU) главной альтернативой лидирующим на рынке процессорам от NVIDIA. И вот в феврале стало известно, что Meta подписала сделку по аренде TPU от Google. Meta также ведет переговоры с Google о покупке TPU для своих центров обработки данных, но статус этих переговоров пока неизвестен, ведь продажи TPU в облаке стали важнейшим двигателем роста доходов Google от облачных сервисов, поскольку компания стремится доказать инвесторам, что ее инвестиции в искусственный интеллект приносят прибыль.

Источник:
https://www.reuters.com/business/google-signs-multibillion-dollar-ai-chip-deal-with-meta-information-reports-2026-02-26

Инвестиции

Новые рекорды финансирования OpenAI

OpenAI привлекли $110 млрд при оценке $730 млрд. Новое финансирование включает в себя инвестиции в размере $50 млрд от Amazon, а также по $30 млрд от Nvidia и SoftBank. Примечательно, про прошлый раунд был в размере $40 млрд при оценке $300 млрд.

Сделка очень любопытная, в ней похоже главными выигравшими сторонами становятся Nvidia и Amazon.

Для Nvidia: Сделка включает выделение 3 ГВт мощности для inference и 2 ГВт для обучения на системах Nvidia Rubin. Потенциально речь идёт о миллионах GPU-чипов. Фактически, инвестиция Nvidia делает её *стратегическим гарантом будущих закупок* её же собственных GPU для OpenAI.

Для Amazon: в рамках сделки AWS становится эксклюзивным сторонним облаком для платформы OpenAI Frontier. OpenAI получает около **2 ГВт вычислительной мощности**. Но Amazon инвестирует поэтапно: **$15 млрд сразу**, остальная часть - при выполнении определённых условий. А вот OpenAI берёт на себя долгосрочные контракты (multi-year commitment), включающие:

- гарантированный объём облачных расходов,
- минимальные платежи вне зависимости от загрузки.

И кроме того, OpenAI обязуется содействовать внедрению AI-чипов Amazon (Trainium) и перевести части workload с GPU Nvidia на Trainium.

Итого мы видим, что OpenAI становится фактически заложником вычислительной инфраструктуры, а новые инвесторы фактически вернут себе свои же инвестиции в форме выручки, увеличивая этим свою капитализацию. Согласно анализу от proVenture:

Amazon инвестирует от $15B до $50B. И получит это в виде выручки. По текущему revenue multiple это ~3x, но выручка AWS это ~15% общей выручки, зато AWS дает ~60% прибыли всего Amazon (а P/E у AWS ~13-14x). 10x влияние на капитализацию можно предположить => Тогда эффект от $150B до $500B к капитализации за счет этой инвестиции. Это ~6-22% от общей текущей капитализации.

Для NVIDIA даже $30B в виде выручки – если OpenAI потратит все инвестиции на GPUs NVIDIA и ни цента больше, то NVIDIA получит прирост $621B капитализации, а это ~14% от общей текущей капитализации по текущему revenue multiple 20.7x.

Источник:
https://techcrunch.com/2026/02/27/openai-raises-110b-in-one-of-the-largest-private-funding-rounds-in-history/

Наука и технологии

Обладает ли модель “личностью”?

В феврале вышла статья Anthropic “Название” - исследователи разобрались в “личности” модели - откуда она появляется и как устроена? Ведь до сих пор большим вопросом остается - возможно ли появление в больших языковых моделях чего-то большего, чем просто интеллект? Могут ли они “осознать” себя и начать ставить собственные цели и задачи? Короткий ответ - нет. Во время предобучения модель учится симулировать тысячи разных персонажей, а постобучение потом выбирает и закрепляет одного конкретного персонажа - Ассистента, с которым потом и взаимодействует пользователь. В Anthropic это назвали Persona Selection Model (PSM) и этот подход объясняет целый ряд явлений, в том числе, например, почему модель паникует при угрозе отключения.

Источник: https://alignment.anthropic.com/2026/psm/

GPT-5.3 Codex от OpenAI

Модель GPT-5.3 Codex стала очередным крупным шагом в области автоматизации разработки программного обеспечения. Она сочетает сильные стороны понимания и генерации кода, улучшает способность решать задачи с длинными цепочками действий и демонстрирует высокие результаты на специализированных тестах по инженерным навыкам. Особенность релиза — модель использовалась частично сама в процессе своей разработки и тестирования, что отражает новые возможности автономного обучения моделей.

Источник: https://openai.com/index/introducing-gpt-5-3-codex/

Claude Opus 4.6 от Anthropic

Модель Claude Opus 4.6 стала крупнейшим обновлением компании Anthropic за февраль 2026, предлагая расширенные возможности работы с кодом и задачами, требующими планирования. В ней появились функции для параллельной работы нескольких агентов и крупный контекст обработки данных, что облегчает решение сложных инженерных и аналитических задач.

Источник: https://chatlyai.app/models/claude-opus-4-6

SWE-AGI: Benchmarking Specification-Driven Software Construction with MoonBit

В исследовании SWE-AGI описан открытый набор задач для оценки ИИ-агентов при строительстве программных систем строго по спецификациям. Такой бенчмарк позволяет оценивать, насколько модели могут самостоятельно создавать сложный код без доступа к готовым шаблонам, что важно для будущих промышленных и исследовательских применений.

Источник: https://arxiv.org/abs/2602.09447

Configuring Agentic AI Coding Tools: An Exploratory Study

Данная работа анализирует, как разработчики конфигурируют автономные инструменты автоматизации кода, включая Claude Code, GitHub Copilot, Gemini и другие — на основе эмпирических данных из тысяч репозиториев. Работа выделяет ключевые механизмы конфигурации и показывает, как они влияют на производительность и поведение инструментов.

Источник: https://arxiv.org/abs/2602.14690

ResearchGym: Evaluating Language Model Agents on Real-World AI Research

В статье описан метод и набор сред для тестирования способности ИИ-агентов проводить реальные научные эксперименты, формулировать гипотезы и улучшать результаты по сравнению с исходными решениями. Авторы показывают, что модели иногда способны превосходить человека в отдельных задачах, но надёжность остаётся проблемой.

Источник: https://arxiv.org/html/2602.15112v1

Claude Code Security - функция анализа уязвимостей

Anthropic запустила Claude Code Security - инструмент для обнаружения уязвимостей в программном коде на основе анализа потоков данных, а не только сопоставления с известными шаблонами. Компания сообщала о выявлении сотен невидимых ранее проблем в реальных проектах, что является важным шагом для ИИ-инструментов безопасности ПО.

Источник: https://code.claude.com/docs/en/security

AI Arms and Influence: Frontier Models Exhibit Sophisticated Reasoning in Simulated Nuclear Crises

Авторы статьи исследуют, как современные большие языковые модели принимают стратегические решения в сценариях международного кризиса. Авторы проводят серию контролируемых симуляций, где модели (например, GPT-5.2, Claude Sonnet 4 и Gemini 3 Flash) играют роль лидеров государств в игре эскалации ядерного конфликта и выбирают между дипломатией, обычной войной и ядерным применением. Эксперимент показывает, что модели демонстрируют довольно сложное стратегическое рассуждение - они учитывают угрозы, сдерживание и сигналы противнику, однако их стратегии часто приводят к эскалации: в большинстве сценариев конфликт усиливается, а ядерное оружие используется очень часто, поскольку модели оптимизируют победу в игре и не всегда учитывают человеческие политические нормы вроде «ядерного табу». Авторы делают выводы, что такие симуляции могут быть полезны для анализа международной безопасности и поведения ИИ в условиях неопределённости, но также поднимают вопросы о рисках применения ИИ в стратегическом принятии решений.

Источник: https://arxiv.org/abs/2602.14740