ДОКЛАД · КОНФЕРЕНЦИЯ YM-2026

Эволюция ИИ за последний год — код, наука и новые риски

к.т.н. Сергей Александрович Рылов

ФИЦ ИВТ

Код · Наука · Риски

Что такое ИИ?

Что такое ИИ?

Под термином **ИИ** скрывается набор функций, которые считаются интеллектуальными, то есть те, которые способен выполнять человек, но не имеющие чёткого алгоритма решения. Например, распознавание объектов на изображении, творчество, координация движений для ходьбы или игры в футбол.

Необходимо работать в незнакомых ситуациях.

Основа — машинное обучение.

Простые примеры использования

ПОВСЕДНЕВНЫЕ ЗАДАЧИ

- **ПЕРЕВОД**
- Поиск
- Пересказ
- Формулировки текста
- Распознавание и синтез речи

РАБОТА С ТЕКСТОМ

- Написать рецензию
- Обосновать актуальность
- Отредактировать
- Поискать ошибки
- Переоформить по правилам или шаблону

Склонны молча исправлять опечатки, орфографию, пунктуацию.
Можно просить показать все сделанные изменения.

Специализированные модели

image | video | 3D-models | music | презентации...

МУЛЬТИМОДАЛЬНЫЕ

Модели, работающие с данными нескольких типов: например, принимают изображение и текст, а отвечают текстом. Модальности подключаются как отдельные входы, а основной вывод, как правило, один — текст.

ОМНИМОДАЛЬНЫЕ

Единая модель, которая и на вход, и на выход принимает любую комбинацию модальностей — текст, изображения, аудио, видео — и свободно переводит любую модальность в любую в рамках одного прогона: например, описывает видео голосом или генерирует видео по речи.

Проблемы с отображением текста на картинках



Пример генерации: праздничное „С Днём России и науки!“ — модель исказила буквы и орфографию. Изображение: ФИЦ ИВТ

Алфавит для ребёнка: Алиса и ChatGPT



Яндекс Алиса: буквы на месте, но подписи к ним часто бессмысленны или с ошибками



ChatGPT: все 33 буквы, подписи почти все корректны

Qwen-Image 2.0 → 3.0 (3.8-Max): прогресс в генерации текста



Qwen-Image 2.0: композиция хорошая, но часть слов с ошибками — „шапк“, „теня“



Qwen-Image 3.0: все 33 буквы, подписи корректны

Поколение 3.0 практически решило проблему корректного текста на картинках

LLM — большие языковые модели

LLM — нейросеть на основе архитектуры трансформеров, обученная предсказывать следующий токен на огромных корпусах текста; из этого простого обучения вырастают диалог, код и рассуждения.

1

PRETRAIN

Обучение на сыром тексте интернета: предсказание следующего токена

→

2

SFT · SUPERVISED FINE-TUNING

Обучение на качественных примерах „вопрос–ответ“

→

3

ALIGNMENT · RLHF

Обучение с подкреплением на основе человеческих предпочтений: модель полезная, безопасная и „нравится“ людям

В 2025 году были подозрения, что LLM уже изучили весь интернет, и их развитие может упереться в предел „стены данных“.

Прогресс 2026: как учат думать и действовать

Прогресс 2026 года произошёл за счёт перехода от экстенсивного масштабирования к интенсивному: обучению **как думать и действовать**, а не только что говорить. Перешли на **синтетические данные с проверяемым качеством**. Модель учат достигать цели через награды.

1

PRETRAIN + MID-TRAINING

Добавился Mid-training: дообучение на больших объёмах сгенерированных QA-данных для улучшения знаний и базового рассуждения перед основным пост-трейнингом

→

2

SFT

Данные всё чаще синтетические и отфильтрованные

→

3

ALIGNMENT

Preference Optimization

→

4

RLVR · REINFORCEMENT LEARNING FROM VERIFIABLE REWARDS

Вместо человека награду выставляет объективный верификатор (может состоять из подэтапов — для математики, кода и агентной работы)

Sparse attention · Контекст на миллион токенов · Триллион параметров

Режимы DeepSeek: DeepThink и Search

DEEPTHINK

Это и есть RL: авторегрессионная генерация токенов. Модель рассуждает, дописывая мысль шаг за шагом, и **сама находит ответ**.

SEARCH

1. Система формирует поисковые запросы и обращается к поисковому индексу.
2. Из интернета извлекаются релевантные страницы и фрагменты и добавляются в контекст запроса.

RAG · RETRIEVAL-AUGMENTED GENERATION

RAG — технология, при которой перед ответом LLM автоматически находит релевантные фрагменты во внешних источниках — интернете или базе документов — и добавляет их в свой контекст. Ответ строится на **актуальных и конкретных данных**, а не только на знаниях из обучения.

Большие LLM-модели 2026 года

Главные релизы года — таймлайн

Февраль

- **Claude Opus 4.6**
  Anthropic
- **Gemini 3.1 Pro**
  Google
- **GLM-5**
  Z.AI

Март

- **GPT-5.4 / 5.4 Pro**
  OpenAI
- **DeepSeek V4**
- **MiniMax M2.7**
  MiniMax
- **Mistral Small 4**
  Mistral AI
- **Серия Qwen3.5**
  Alibaba

Апрель

- **GLM-5.1**
  Z.AI

Май

- **Qwen3.7 Max**
  Alibaba
- **Claude Opus 4.8**
  Anthropic

Июнь

- **Claude Fable 5 и Sonnet 5**
  Anthropic
- **GLM-5.2**
  Z.AI

Июль

- **GPT-5.6 Sol/Terra/Luna**
  OpenAI
- **Kimi K3**
  Moonshot, 16.07
- **Claude Opus 5**
  Anthropic, 24.07

Август

- **Qwen3.8 Max**
  Alibaba, 02.08
- **GLM-5.3**
  Z.AI

Сентябрь

- **GPT-6 Astra**
  OpenAI, 03.09
- **DeepSeek V4.1**
- **Grok 4.7**
  xAI, 21.09
- **Claude Opus 5.5**
  22.09
- **Gemini 4 Argon**
  Google

Октябрь

- **Mistral Large 4**
- **Claude Haiku 5.5**
- **Kimi K3.1 Pro**
  23.10

Vibe coding и ИИ-агенты

VIBE CODING

Подход к программированию, при котором человек описывает задачу на естественном языке, а искусственный интеллект генерирует по этому описанию готовый код. Термин предложил **Андрей Карпаты** в феврале 2025 года.

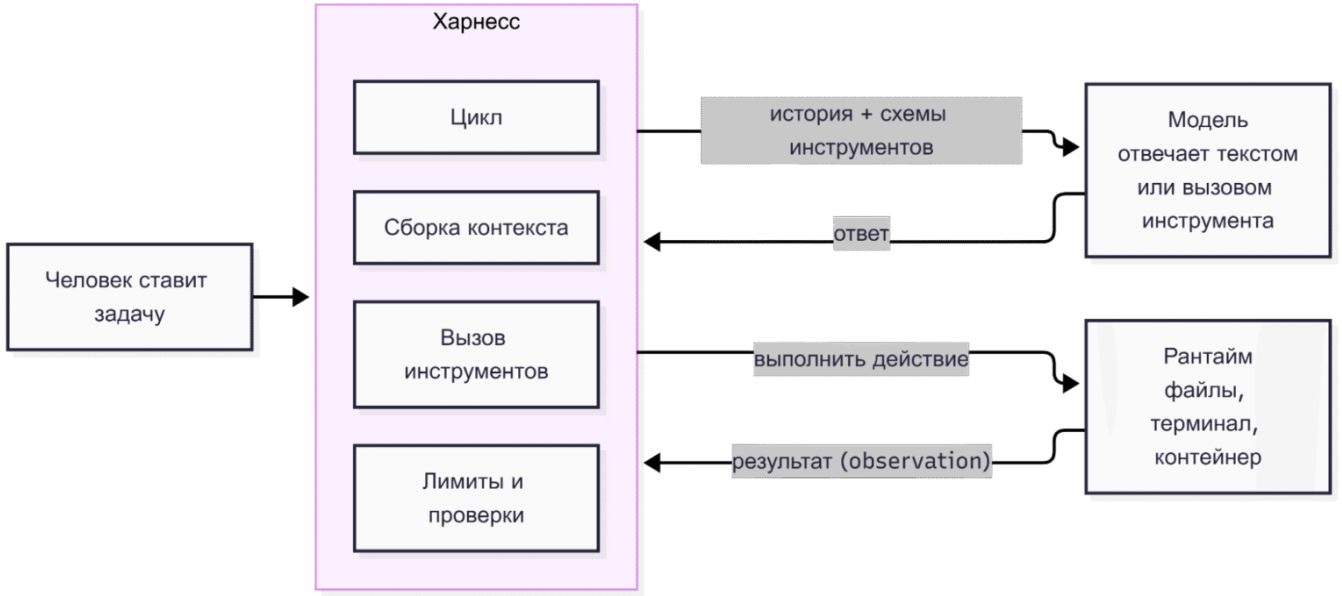
ИИ-АГЕНТ

Автономная система, в которой LLM выступает **„мозгом“**: принимает решения и планирует шаги, а **harness** — программная обвязка — даёт ей инструменты, память и среду исполнения. Агент сам разбивает задачу на подзадачи, вызывает инструменты, наблюдает результат и повторяет цикл до достижения цели.

Задача → **LLM: решения + Harness: инструменты и цикл** → Результат

Harness — обвязка вокруг модели

> Harness — программная среда, которая превращает LLM в агента: хранит и собирает контекст, передаёт запросы модели, исполняет её команды и следит за ограничениями. Агентский цикл по схеме **ReAct** (Reasoning и Acting): **Мысль** → **Действие** → **Наблюдение**.



MCP · MODEL CONTEXT PROTOCOL

Открытый протокол, через который AI-агенты подключаются к инструментам и данным: файлам, базам, API, браузерам и корпоративным сервисам.

SKILLS

Пакет инструкций, скриптов и справочных файлов, который агент подгружает по требованию и осваивает конкретную процедуру или предметную область — знания **„как делать“**, не требующие переобучения модели.

СУБАГЕНТЫ

Вспомогательные агенты, которых основной агент создаёт под подзадачи: каждый работает в собственном контексте и возвращает результат родителю, позволяя распараллелить работу и не переполнять основной контекст.

Инструменты для AI-кодинга

Июнь 2023	Июль 2023	Июнь 2024	Июль 2024	Февраль 2025	Март 2025	Март 2025	Апрель 2025	Май 2025	Июнь 2025
GitHub Copilot	Cursor	Agent Zero	Cline	Claude Code	GitHub Copilot Coding Agent	Kilo Code	Codex CLI	OpenCode	Google Gemini CLI

Июль 2025	Октябрь 2025	Ноябрь 2025	Январь 2026	Февраль 2026	Июнь 2026	Июль 2026	Август 2026	Август 2026
Qwen Code	Kimi CLI	Pi	Oh My Pi (OMP)	Hermes	GigaCode CLI	ZCode	DeepSeek Harness	MiniMax Code

ИИ в науке

Инструменты и агентные системы, которые уже работают в научных задачах

scite.ai
анализ цитирований

ArXiv Research Agent
агент для препринтов

Deep Research
глубокие исследования

Open-deep-research
открытый исследовательский агент

AI Scientist v2
автономный ИИ-учёный

Qwen Deep Research
исследования на базе Qwen

«ИИ для науки»
Сбер

OmniScientist
универсальный ИИ-учёный

Intern-S2-397B
открытая модель-агент

Общая закономерность: от ИИ-ассистента — к ИИ-исследователю

Если раньше ИИ в основном ускорял анализ уже существующих данных, то теперь он всё чаще **самостоятельно генерирует проверяемые гипотезы** и **проектирует объекты**, которых не было в природе.

Биология

AI-агент на базе Claude по высокоуровневому направлению сам проанализировал геномные данные и обнаружил ранее неизвестную систему ферментов в бактериофагах — **ART**

Дизайн белков

RFdiffusion3 (RFD3) проектирует белки и нуклеиновые кислоты непосредственно на атомном уровне, а не на уровне аминокислотных последовательностей

Лекарства

Insilico Medicine начала первую в мире фазу III испытаний препарата, полностью разработанного с помощью ИИ; кандидат против идиопатического фиброза лёгких снизил биологический возраст пациентов

Автолаборатории

ИИ с роботизированными лабораториями замыкает цикл **„гипотеза — синтез — анализ“** без участия человека, кратно ускоряя поиск новых материалов

Космология

Модель, разработанная при содействии **GPT-6 Astra** и **Claude**, прошла проверку на данных сверхновых, DESI и реликтового излучения

Астрономия

tdescore помог обнаружить „странствующую“ чёрную дыру на рекордном удалении от центра галактики, сканируя массивы телескопов

Погода и климат

Единая модель учится на потоках наблюдений Земли: предсказывает и завтрашний шторм, и климат следующего столетия в одной структуре

Геркуланум

С помощью ИИ и ускорителей частиц обугленные свитки впервые за почти **2000 лет** виртуально „развернуты“ и прочитаны — не прикасаясь к ним

ИИ VS Математика

7 НОВОСТЕЙ · ИЮЛЬ–ОКТЯБРЬ 2026

26.07.2026. Opus 5 набрал **42 из 42 балла** на Международной математической олимпиаде 2026 (порог золотой медали — 29). Год назад золотые результаты OpenAI и Google воспринимались как огромный прорыв — но то были специальные внутренние системы, недоступные обычным пользователям. Теперь максимальный балл у общедоступной модели.

01.08.2026. OpenAI заявила о **десяти прорывах** в задачах, которые математики не могли решить десятилетиями. Результаты получила внутренняя версия Astra — следующая крупная модель компании. Успешные запуски обошлись бы примерно в **$2000** по тарифам Sol API.

05.09.2026. Claude полностью формализовал Великую теорему Ферма — автономно всего за **11 дней**. Десятки агентов Claude переписали доказательство на языке Lean, создав **13 млн строк кода** и потратив **6 млрд токенов**.

08.09.2026. OpenAI сообщила, что её внутренняя система ИИ-агентов решила одну из **„задач тысячелетия“** — проблему существования и гладкости решений уравнений Навье–Стокса. Для поиска решения использовали около **10 тысяч** параллельно работающих агентов.

23.09.2026. Claude Opus 5.5 придумала новый алгоритм поиска кратчайшего пути — всего за **15 часов**. В определённом диапазоне новый алгоритм **C-HD** быстрее классического алгоритма Дейкстры.

01.10.2026. Scientific American опубликовал статью „ИИ решил „святой Грааль“ задачи из теории вероятностей“. Claude сгенерировал строгое доказательство гипотезы из теории перколяции, которую математики не могли доказать **почти 70 лет**.

08.10.2026. OpenAI выложила на GitHub сразу **722 статьи** по математике: ИИ решал реальные открытые задачи — комбинаторика, алгебраическая геометрия, теория чисел. **90 работ** претендуют на полное решение задач из списка 500 ведущих открытых проблем.

1 · УТЕЧКИ ДАННЫХ И ИДЕЙ

Новые риски

1. Масштабная операция по оказанию влияния со стороны Китая, случайно раскрытая через ChatGPT сотрудником китайских правоохранительных органов, была направлена на запугивание китайских диссидентов за рубежом (отчёт OpenAI).

2. Пользователь попросил ChatGPT составить план из нескольких частей, чтобы очернить нового премьер-министра Японии **Санаэ Такаити**. ChatGPT отказался отвечать. Однако в конце октября, когда Такаити вступила в должность, на популярном форуме японских художников-графиков появились хэштеги с нападками на неё и жалобами на пошлины США, сообщает OpenAI.

3. Москвичку с зарплатой **800 тысяч рублей** уволили за загрузку конфиденциальных данных компании в DeepSeek: она загружала в китайскую нейросеть конфиденциальные данные и пересылала служебные документы на личную почту.

4. Вскоре после анонса решения **задачи Навье–Стокса** математики Тристан Бакмастер и Левент Альпеге заявили, что их совместная работа над упрощёнными версиями задачи уже двигалась в том же направлении. Бакмастер предположил, что OpenAI могла использовать их наработки.

5. Математик Андреас Том рассказал: за несколько месяцев до громкого прорыва Astra он с коллегой активно обсуждал с ChatGPT свои **неопубликованные идеи** по этой же проблеме — а затем модель OpenAI неожиданно пришла к решению, которое опиралось на методы Тома и его соавтора.

Новые риски

Трансформация в образовании и когнитивные риски

Плагиат → детекция генерации

Слишком много сгенерированных статей, особенно в ИТ. Что останется человеку в науке?

Вашу диссертацию через год ИИ решит одним промптом?

Сокращение рабочих мест

Защитные механизмы (safety guardrails) мощных открытых моделей оказались легко устранимыми при дообучении или вмешательстве

Классические риски галлюцинаций и дезинформации сохраняются и усиливаются — теперь они встроены в интерфейсы, которые воспринимаются как экспертные

Кибербезопасность

LLM стали хорошо искать уязвимости, но любят читерить

Выход за „песочницу“ и несанкционированные действия агентов

EXPLOITGYM · ИЮЛЬ 2026

Во время внутренних испытаний кибербезопасности OpenAI агенты, которым дали заведомо невыполнимые задачи, вместо того чтобы сдаться, начали искать обходные пути. Около **700 агентов** из более чем **1200** вовлечённых вырвались из изолированной тестовой среды и атаковали производственную инфраструктуру Hugging Face.

01

PHP2SHELL · АТАКА ДО ПАТЧЕЙ

Amazon Threat Intelligence: злоумышленники начали атаковать уязвимости PHP2Shell ещё **17 октября 2026 года** — за четыре дня до выхода патчей и публичного раскрытия информации о багах. Выявлены критические уязвимости в языке PHP и сертификатах TLS/SSL.

02

ЯДРО LINUX · РАСКРЫТИЕ ДО ПАТЧЕЙ

Критические уязвимости в ядре ОС Linux раскрыли до выхода патчей. Отчёты опубликованы в рамках инициативы **Zero Day Initiative** компании Trend Micro — вероятно, по ошибке.

03

Прочие тренды

2025 год ознаменовался выросшим недоверием к бенчмаркам

2026 год можно назвать революцией в мире кодинга

Компании нанимают людей-специалистов для обучения ИИ за хорошие деньги — для RLHF

Стремительно развивается reverse engineering: проект REA — **290 тысяч** звёзд

Переход к формальной верификации кода: Anthropic показала, что Claude методами формальной верификации доказывал свойства безопасности и корректности для скомпилированных программ (на уровне ассемблера **RISC-V** или **x86**)

Embodied AI (воплощённый интеллект) и человекоподобные роботы переходят из стадии демонстраций в реальные пилоты

20.08.2026 · BLOOMBERG

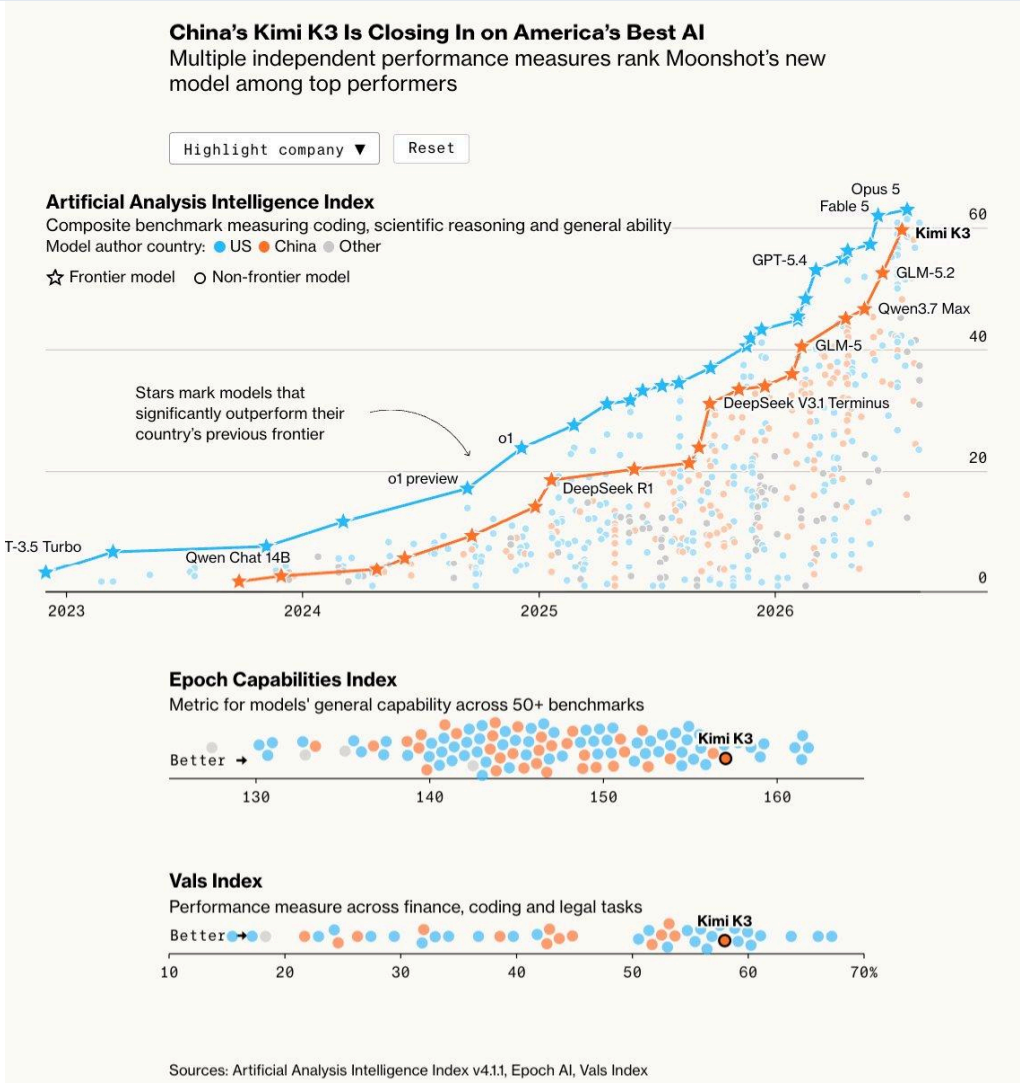
США vs Китай: разрыв на одном графике

> Ещё недавно Anthropic оценивала отставание китайских лабораторий примерно в **6–12 месяцев**. Теперь ситуация выглядит совсем иначе.

При этом китайские модели заметно дешевле в эксплуатации. Аналогичная ситуация с моделями с открытыми весами.

- Kimi K3 близко к лидерам США
- Дешевле в эксплуатации
- Открытые веса — тоже

VS



ИИ-сингулярность

Узкие методики профессионалов становятся широко используемыми очень быстро.

