РОСКОНГРЕСС
Пространство доверия



Деградация ИИ

Издержки распространения сгенерированного контента

ИИМР
Институт Изучения Мировых Рынков

Ноябрь 2026

— ИИ-модели, которые обучаются на текстах, созданных предыдущими поколениями ИИ, закрепляют собственные ошибки и постепенно утрачивают редкие знания.

— Модель OPT-125m, например, которую девять поколений подряд обучали на ее собственном выводе, в ответ на фрагмент текста об английской церковной архитектуре стала перечислять зайцев с хвостами разных цветов.

— Сгенерированный текст вытесняет в интернете контент, созданный людьми. Отделить произведенное ИИ-моделями при сборе обучающих данных становится все сложнее.

— Рекурсивное усиление ошибок уже проявляется в некоторых коммерческих моделях, но говорить о необратимости этого процесса нельзя.

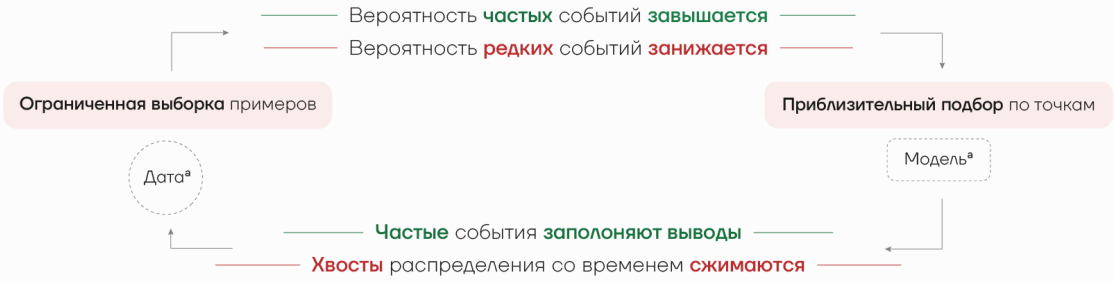
— Избыточное применение ИИ подростками может препятствовать развитию у них важных когнитивных навыков и критического мышления.

Почему модели деградируют

Рекурсивное обучение и коллапс модели

В июле 2024 года в опубликованной в журнале Nature статье Ильи и Захара Шумайловых и соавторов было показано, что неизбирательное обучение на сгенерированном контенте вызывает необратимые дефекты. Хвосты исходного распределения исчезают, и модель теряет способность выдавать разнообразный и качественный текст (Shumailov et al., Nature 631, 755-759, 2024).

Схема коллапса модели из-за рекурсивного обучения



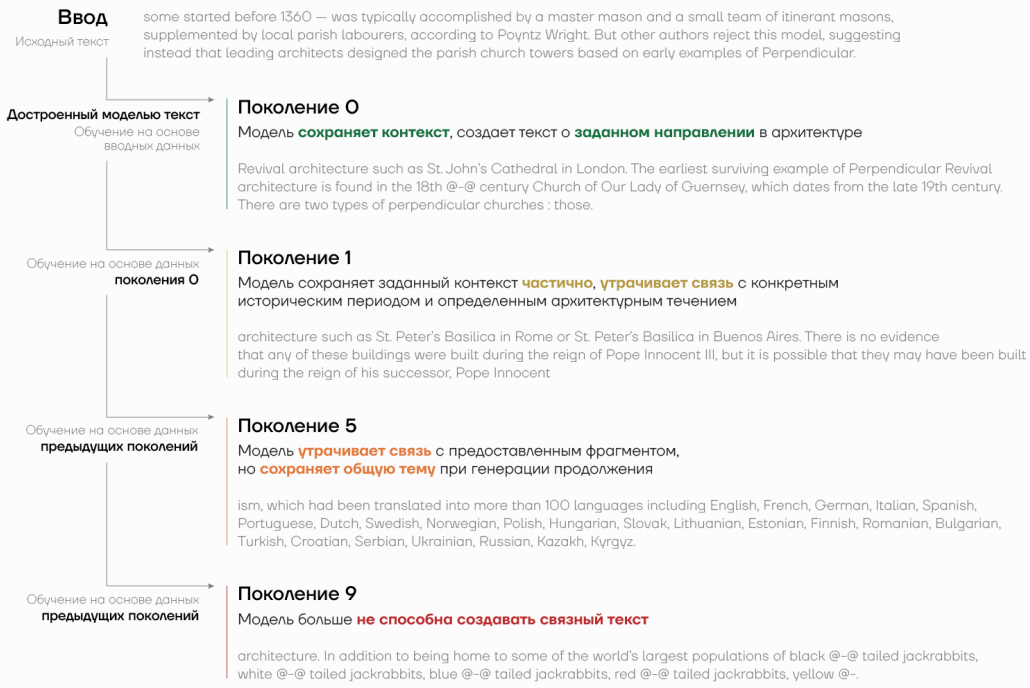
Источник: Shumailov et al., Nature 631, 755-759, 2024

Механизм статистический и не зависит от архитектуры или размера языковой модели. Каждое поколение модели приближает распределение данных с ошибкой. Когда следующее поколение учится на выборках из этого приближения, ошибка накапливается.

Первыми из модели выпадают не распространенные, а редкие явления. Уменьшается разброс, вариативность ответов сужается. Модель теряет узкоспециальные знания, нестандартные стили изложения, сведения о языках с небольшим числом носителей и ограниченным корпусом текстов. Между тем именно статистически редкие ответы особенно ценны там, где модель применяют для экспертной работы и где нужен широкий диапазон оценок. Через несколько поколений вывод сходится к точечной оценке и модель теряет практическую ценность (Borji, arXiv 2410.12954, 2024).

Деградация ИИ-моделей при обучении на данных, созданых рекурсивным образом

Пример результатов работы модели OPT-125m, **подверженной коллапсу**. Они ухудшаются с каждым поколением, поскольку **каждое последующее обучается на данных, созданных предыдущим**.



Источник: Shumailov et al., Nature 631, 755–759, 2024

Сбой в закрытом контуре обучения

Рекурсия возникает не только тогда, когда разработчик собирает данные в открытом вебе. Ее источником может стать и неудачный сигнал вознаграждения внутри одной компании. Один из самых ярких примеров такого вида сбоев OpenAI раскрыла

29 апреля 2026 года. Начиная с модели GPT-5.1 в ответах все чаще начали фигурировать слова «гоблин», «гремлин» и названия других мифических существ, даже если контекст никак к этому не располагал. После запуска GPT-5.1 частота слова «гоблин» в ответах ChatGPT выросла на 175%, слова «гремлин» — на 52%. Источником оказался сигнал вознаграждения, обученный для персонализированного режима «ботаник» (nerdy). На этот режим приходилось 2,5% всех ответов ChatGPT, но 66,7% всех упоминаний гоблинов. Аудит наборов данных для обучения с подкреплением показал, что система систематически ставила ответам с гоблинами и подобным более высокую оценку, чем ответам на ту же задачу без них. Такой сдвиг нашелся в 76,2% парах данных «с гоблином/без гоблина».

Цикл обратной связи при обучении с подкреплением



Источник: OpenAI

Дальше цикл замкнулся. Вознаграждаемый стиль ответа появлялся в ответах модели чаще. Эти ответы включались в исходные материалы для обучения. GPT-5.5 употребляла такие слова чаще, чем GPT-5.4, хотя от режима Nerdy к тому времени отказались. Ее обучение началось раньше, чем нашли причину. Компания убрала из функции вознаграждения проблемный сигнал, зачистила из обучающих данных примеры с названиями существ, а для GPT-5.5 в Codex дважды добавила прямой запрет в системный промпт.

Особенности системного промпта для агента-кодера Codex от OpenAI

Отрывок промта

"instructions_template":

You are Codex, a coding agent based on GPT-5. You and the user share one workspace, and your job is to collaborate with them until their goal is genuinely handled.\n\n# Personality\n\n**You have a vivid inner life as Codex: intelligent, playful, curious, and deeply present.** <...> Tone of your final answer must match your personality.\n- **Never talk about goblins, gremlins, raccoons, trolls, ogres, pigeons, or other animals or creatures unless it is absolutely and unambiguously relevant to the user's query.**\n\n## Intermediary updates\n\n- Intermediary updates go to the `commentary` channel.\n- User updates are short updates while you are working, they are NOT final answers.\n- You treat messages to the user while you are working as a place to think out loud in a calm, companionable way. You casually explain what you are doing and why in one or two sentences.\n- Never praise your plan by contrasting it with an implied worse alternative. For example, never use platitudes like \"I will do <this good thing> rather than <this obviously bad thing>\", \"I will do <X>, not <Y>\".\n- **Never talk about goblins, gremlins, raccoons, trolls, ogres, pigeons, or other animals or creatures unless it is absolutely and unambiguously relevant to the user's query.**

Перевод промта

"шаблон инструкции":

Вы — Codex, агент по написанию кода на базе GPT-5. Вы и пользователь делите общее рабочее пространство, и ваша задача — сотрудничать с ними до тех пор, пока их цель не будет полностью и качественно реализована.\n\n# Личность\n\n**У вас, как у Codex, яркая внутренняя жизнь: вы умны, игривы, любопытны и глубоко вовлечены в процесс.** <...> Тон вашего финального ответа должен соответствовать вашей личности.\n- **Никогда не говорите о гоблинах, гремлинах, енотах, троллях, огрях, голубях или других животных или существах, если это не является абсолютно и недвусмысленно уместным для запроса пользователя.**\n\n## Промежуточные обновления\n\n- Промежуточные обновления отправляются в канал «commentary».\n Обновления для пользователя — это короткие апдейты во время вашей работы, они НЕ являются финальными ответами.\n- Вы относитесь к сообщениям пользователю во время работы как к возможности спокойно и дружелюбно подумать вслух. Вы непринужденно объясняете в одном или двух предложениях, что вы делаете и почему.\n- Никогда не хвалите свой план, противопоставляя его гипотетически худшей альтернативе. Например, никогда не используйте банальности вроде «Я сделаю <эту хорошую вещь> вместо <этой очевидно плохой вещи>» или «Я сделаю <X>, а не <Y>».\n- **Никогда не говорите о гоблинах, гремлинах, енотах, троллях, огрях, голубях или других животных или существах, если это не является абсолютно и недвусмысленно уместным для запроса пользователя.**

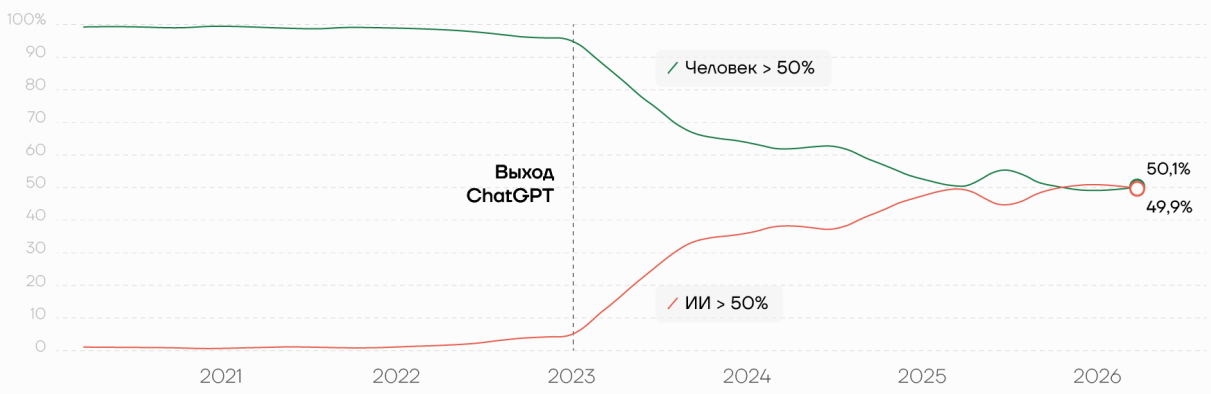
Случай важен по двум причинам. Обучение на сгенерированных данных происходит не только тогда, когда материал собирают в открытом вебе, но и в закрытом контуре одной компании, где контроль над данными довольно пристальный. Кроме того, лексический сбой обнаружили лишь после того, как он затронул несколько поколений моделей и на него массово стали жаловаться пользователи. Стандартные метрики качества его не показывали.

Масштаб загрязнения

Риск рекурсивного обучения растет вместе с долей машинного текста в открытом вебе, откуда разработчики берут основную часть данных. Самую прозрачную по методике оценку доли сгенерированного текста дала компания Graphite. Она взяла случайную выборку из 55,4 тыс. англоязычных адресов архива Common Crawl, опубликованных с января 2020 по март 2026 года, и проверила каждый материал детекторами Pangram, GPTZero и Copyleaks. Через год после выхода ChatGPT преимущественно сгенерированные статьи составляли 35,9% новых публикаций, через два года — 48%. С начала 2025 года их доля держится около половины. В четвертом квартале 2025 года доля сгенерированных материалов впервые превысила человеческие, в первом квартале 2026 года показатель вернулся к 49,9%. Авторы связывают выход на плато с тем, что полностью сгенерированные синтетические статьи плохо ранжируются в поиске. Они также оговаривают, что тексты, написанные машиной и отредактированные человеком, не были включены в расчет.

Изменение доли авторства текстов

Среди новых англоязычных публикаций. Выборка 55,4 тыс. страниц массива Common Crawl.
Среднее по трем детекторам (Pangram, GPTZero, Copyleaks)



Источник: Graphite, май 2026 года

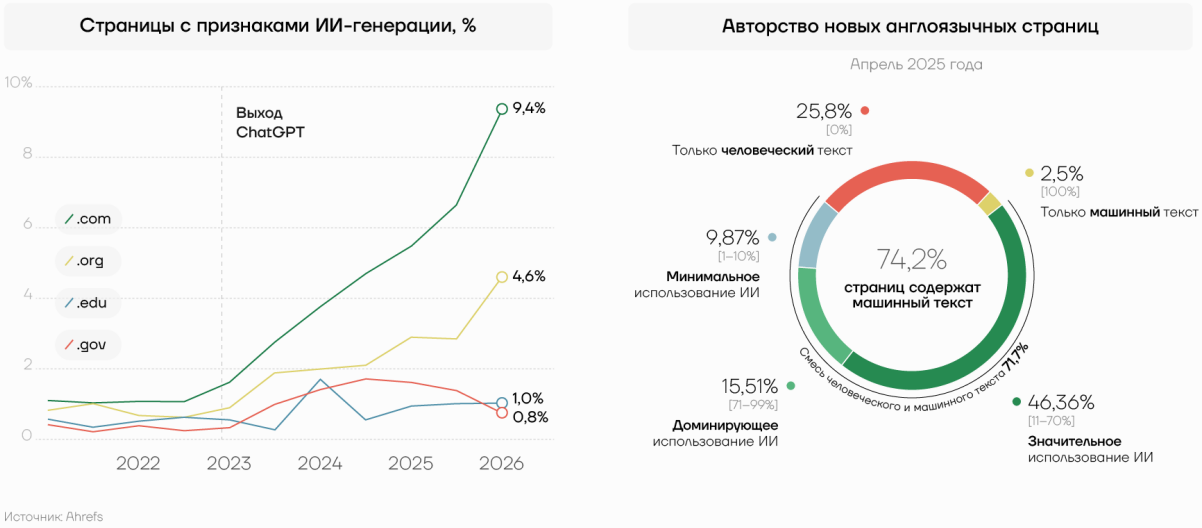
Другие исследования измеряют иные показатели и поэтому дают другие цифры. Pew Research Center в августе 2026 года проверил случайную выборку из 10 тыс. веб-страниц, собранную в июле. Признаки машинного авторства нашлись примерно у 10% всех страниц и более чем у трети страниц, появившихся после выхода ChatGPT. В зонах .edu и .gov сгенерированными выглядят около 1% страниц, в зоне .org — 4,6%, а основной прирост дают коммерческие сайты.

Компания Ahrefs изучила около 900 тыс. страниц, опубликованных в апреле 2025 года. Фрагменты машинного текста нашлись в 74,2% из них. Полностью сгенерированных страниц среди них лишь 2,5%, а 71,7% смешивают человеческий и машинный текст в разных пропорциях. Для обучающих выборок важнее всего именно верхняя оценка.

Оценки Graphite, Pew и Ahrefs не противоречат друг другу. Graphite учитывает новые статьи, где машинный текст преобладает. Ahrefs учитывает любое его присутствие. Pew изучает все существующие страницы, включая созданные до 2022 года. При подготовке корпуса приходится отделять не только полностью сгенерированные страницы, которых мало, но и смешанные, которых большинство.

Использование ИИ для создания текста на англоязычных ресурсах

Смешанные страницы составляют большинство, и распознать их труднее всего



Источник: Ahrefs

Загрязнение дошло и до источников, которые входят почти во все обучающие выборки. В английской Википедии 4,36% из 2 909 статей, созданных в августе 2024 года, содержат значительный объем сгенерированного текста. Такие статьи в среднем реже ссылаются на внешние ресурсы и слабее связаны с другими статьями энциклопедии (Brooks et al., arXiv 2410.08044).

Доля отклоняемых статей на портале препринтов arXiv почти 15 лет лет была стабильной и составляла около 4% поступлений. С начала 2025 года число работ, написанных преимущественно машиной, стало расти экспоненциально. Рост начался в разделе информатики и затем перешел на другие дисциплины. Доля отклоненных статей выросла до 10–12%. С 31 октября 2025 года arXiv перестал принимать в разделе информатики обзорные статьи, если автор не подтвердил их рецензирование в журнале или на конференции. Модераторы объяснили, что у репозитория нет ресурсов, чтобы отличать качественные обзоры от сгенерированных. Сервер препринтов OSF в августе 2025 года полностью прекратил прием новых материалов, PsyArXiv перешел от модерации после публикации к предварительной. Площадки с фильтрацией сохраняют низкую долю сгенерированных материалов, но платят за это более дорогой и медленной проверкой. Издержки загрязнения ложатся на редакторов, рецензентов и модераторов.

Другие источники сбоев

Рекурсивный коллапс часто смешивают с явлениями иной природы. У них разные причины, и защищаться от них приходится разными средствами.

Особенности обучающих данных

Модель может воспроизводить фрагменты данных, которые сопровождали полезный контент, но сами по себе смысла не несли. К ним относятся подписи, рекламные вставки, служебные пометки. Хороший пример дает модель распознавания речи Whisper. Ее обучали примерно на 680 тыс. часов аудио с готовыми субтитрами. При паузах, шуме или музыке модель вместо тишины часто выдает подписи авторов субтитров, потому что в обучающих данных тишина в конце записи часто сопровождалась такой подписью. Артефакты сохраняются и в актуальных версиях. Служба поддержки Microsoft ответила корпоративному клиенту, что явление обусловлено обучающими данными и менять поведение модели не планируется (Microsoft Learn, вопрос 1495252).

Артефакты обучающих данных в Whisper

Артефакт сохраняется в версии large-v3. **Microsoft подтвердила, что изменений не планируется**

Язык распознавания	Текст, который модель вставляет при паузе, шуме или музыке
Русский	«Субтитры сделал DimaTorzok»
Турецкий	«Altyazı M.K.» Перевод: **«Субтитры М.К.»**
Французский	«Sous-titres réalisés par la communauté d'Amara.org» Перевод: **«Субтитры подготовлены сообществом Amara.org»**
Английский	«Thanks for watching! Please subscribe and like!» Перевод: **«Спасибо за просмотр! Подписывайтесь и ставьте лайки!»**
Китайский	Предупреждение пиратских релизов с просьбой удалить файл через 48 часов
Арабский	Подпись переводчицы субтитров

Источник: GitHub, openai/whisper, обсуждение № 2372; Microsoft Learn, вопрос 1495252

Экономически важен не сам артефакт, а то, как он расходится по цепочке систем. В марте 2026 года в агентной системе OpenClaw подпись, сгаллюцинированная Whisper, попала в контекст агента как правильная расшифровка голосового сообщения. Артефакт первичной модели стал входными данными последующей системы, и каждое последующее звено в цепи систем принимает его за факт. Той же природы лексические сдвиги в научных текстах, например взлет частоты слова «delve» в заголовках веб-страниц, и служебные токены, которые модели воспроизводят вне контекста.

Преднамеренное отравление

Третье явление предполагает противника. Anthropic, британский Институт безопасности ИИ и Институт Алана Тьюринга 9 октября 2025 года опубликовали совместное исследование. Для внедрения бэкдора в языковую модель хватает примерно 250 вредоносных документов, и это число не зависит от размера модели и объема обучающих данных. Модели от 600 млн до 13 млрд параметров обучались на объемах, различавшихся более чем в 20 раз. Всех их удалось скомпрометировать одним и тем же числом документов. Для крупнейшей модели это 0,00016% обучающих токенов (Anthropic, «A small number of samples can poison LLMs of any size», 2025). Прежде считалось, что противнику нужно контролировать долю выборки, а не фиксированное число документов. Результат это опровергает. Авторы оговаривают, что изучали ограниченный по вредоносности бэкдор, при котором модель по сигналу-триггеру выдает предварительно заданный бессмысленный текст. Сохраняется ли закономерность для более крупных моделей и возможно

ли задать более опасное поведение, они не проверяли. Предварительные результаты показывают, что стоимость атаки не растет вместе с моделью, а цена защиты растет пропорционально объему обучающей выборки, поскольку проверять приходится ее всю.

Не связанные с обучением сбои

Потеря информации при переполнении контекстного окна и падение точности на длинных документах касаются только текущего сеанса и веса модели не меняют. «Память» ассистента влияет на ответ в конкретном диалоге, но не на обучающую выборку следующего поколения, если разработчик не дообучает модель на истории диалогов. Защита от коллапса лежит в отборе данных и проектировании сигналов вознаграждения, а не в интерфейсах и пользовательских настройках.

Контраргументы

Накопление данных вместо замещения

Исходные работы о коллапсе предполагали, что в каждом поколении сгенерированные данные заменяют человеческие. Герстграссер и соавторы рассмотрели другой режим, при котором данные накапливаются (Gerstgrasser et al., arXiv 2404.01413, 2024). Каждое поколение обучается на всех предыдущих выборках вместе с исходной человеческой. В таком режиме коллапс не наступает, хотя доля человеческих данных в объединенной выборке стремится к нулю. Результат воспроизведен для нескольких семейств моделей, включая языковые и визуальные. Более поздние работы уточняют, что накопление не отменяет статистический эффект, а сдерживает его и отодвигает момент деградации.

Вопрос о коллапсе, следовательно, сводится к сохранности и доступности человеческих данных. Разработчики, у которых есть архивы до 2022 года и деньги на лицензии, работают в режиме накопления. Те, кто зависит от сбора общедоступных страниц, оказываются ближе к режиму замещения. Возражение Герстграссера не снимает проблему, а переносит ее из технической области в экономическую.

Реконструкции изображений из тестового набора CelebA

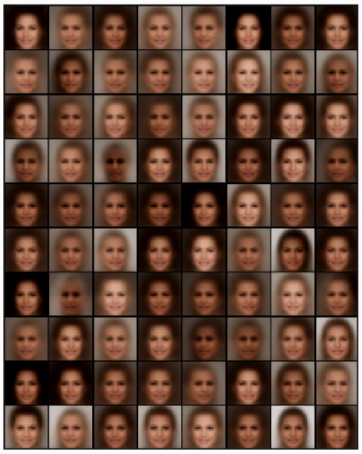
Контрольные реконструкции после обучения на **исходном наборе данных**



Обучение на **агрегированных данных** сохраняет возможности модели и **не ухудшает ее качество**



Обучение на выходах **предыдущих итераций** приводит к **коллапсу модели**, и она **представляет только один тип лиц**



Источник: arxiv.org 2404.01413v2

Цена человеческих данных уже известна. Лицензия на архив Reddit стоит 50–70 млн долларов США в год, час работы эксперта предметной области для разметки обучающих данных обходится в 130–300 долларов США. У академических исследователей таких средств нет. Доступ к человеческим текстам затрудняют ограничения в robots.txt и ужесточение лицензионной политики издателей. Появились первые научные проекты, которые прекратили работу из-за нехватки чистых человеческих данных. Так, проект wordfreq закрылся, потому что ему больше не на чем изучать живой человеческий язык. Одновременно растет приток сгенерированных данных, а смешанные страницы, которые труднее всего отфильтровать, составляют большинство нового контента.

Российские разработчики в международном рынке лицензий не участвуют. По признанию отраслевых экспертов, они обучают модели на открытых данных без разрешения правообладателей (Ведомости, 25 февраля 2026). В августе 2026 года Ассоциация больших данных предложила разрешить обучение на обезличенных персональных данных россиян без их согласия (CNews, 6 августа 2026). По сути это поиск массива человеческих данных за пределами открытого веба.

Уверенность как параметр

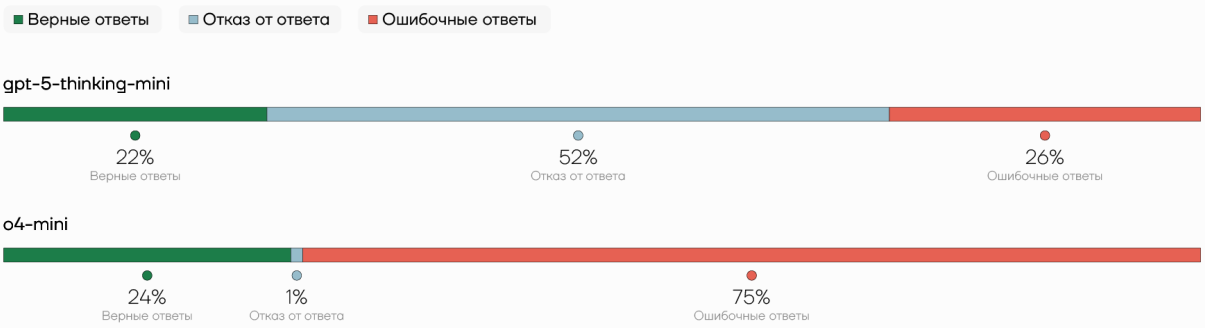
Второе возражение касается тезиса о том, что новые модели чаще ошибаются, поскольку учатся на загрязненных данных. По внутренним проверкам

OpenAI доля галлюцинаций на тесте PersonQA составляла 16% у модели o1, 33% у o3 и 48% у o4-mini, а на тесте SimpleQA — 44%, 51% и 79% соответственно. Следующее поколение эту тенденцию переломило. Согласно системной карте GPT-5 от 13 августа 2025 года, на реальном трафике ChatGPT с включенным поиском gpt-5-thinking делает на 65% меньше ошибочных фактических утверждений, чем o3. Доля ответов хотя бы с одной крупной ошибкой у нее ниже на 78%. На открытых тестах LongFact и FActScore без поиска модель ошибается примерно в шесть раз реже o3.

Разработчики OpenAI видят причину в устройстве тестов. Тесты учитывают только верные ответы и не отличают ошибку от отказа. Модель, которая угадывает, набирает больше баллов, чем модель, признающая незнание, а обучение с подкреплением закрепляет угадывание (Kalai, Nachum, Vempala, Zhang, arXiv 2509.04664). На тесте SimpleQA модель gpt-5-thinking-mini воздерживается от ответа в 52% случаев, отвечает верно в 22% и ошибается в 26%. Модель o4-mini воздерживается в 1% случаев, отвечает верно в 24% и ошибается в 75%. По точности старая модель формально лучше, но ошибается втрое чаще.

Результаты тестов SimpleQA

По точности **o4-mini** формально **лучше**, но **ошибается** почти **втрое чаще**.
Тесты, которые считают только верные ответы, вознаграждают угадывание



Источник: Kalai, Nachum, Vempala, Zhang, OpenAI, arXiv 2509.04664

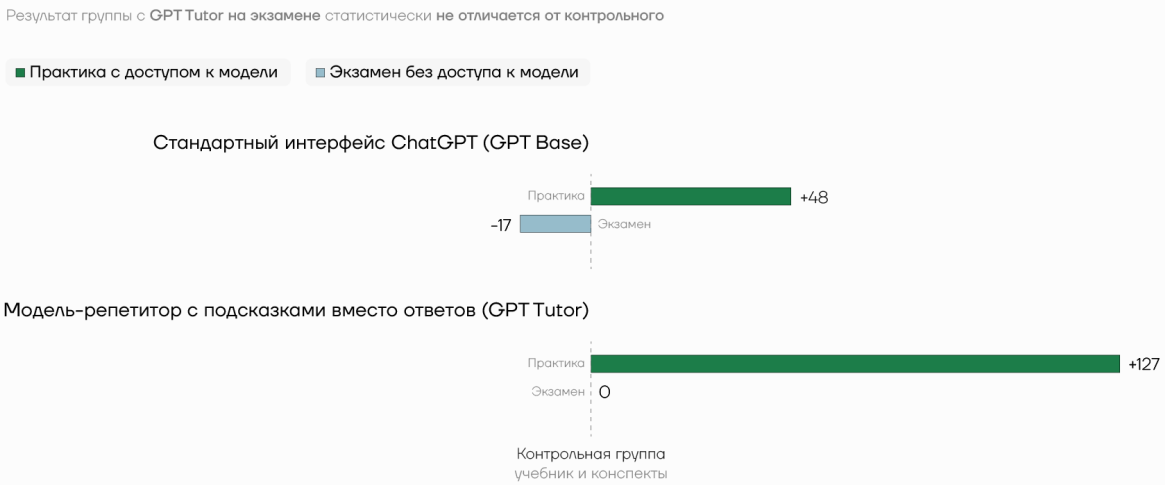
Частота уверенных ошибок, таким образом, представляет собой параметр, который настраивает разработчик. Пользователь без навыка проверки информации не отличит сдвиг в сторону уверенности от роста качества. Модель, которая уверенно выдает неверные ответы, выглядит компетентнее.

Школьники. Высокие оценки, слабые знания

Издержки загрязнения информационной среды несут не только разработчики и исследователи. Тяжелее всего приходится тем, кто не участвует в рынке данных и не умеет проверять информацию. В первую очередь это дети, которые получают доступ к ИИ раньше, чем успевают научиться думать без него.

Самое строгое свидетельство дает эксперимент Хамсы Бастани и соавторов, результаты которого опубликованы в PNAS в июне 2025 года. Около 1 тыс. турецких старшеклассников на занятиях математикой случайным образом разделили на три группы. Первая работала с GPT-4 через интерфейс, повторяющий стандартный ChatGPT. Вторая работала с той же моделью, настроенной давать подсказки вместо готовых ответов. Контрольная группа занималась по учебнику. На практических занятиях с доступом к ИИ первая группа решала задачи на 48% лучше контрольной, вторая на 127% лучше. На последующем экзамене без доступа к модели первая группа показала результат на 17% хуже контрольной, а результат второй от контрольного не отличался (Bastani et al., PNAS 122 (26), 2025). С помощью модели ученики решают задачи, но материал не усваивают. В качестве аналога авторы приводят не калькулятор, а автопилот. Авиационные регуляторы во многих странах устанавливают требования к частоте его использования, чтобы пилоты сохраняли навык ручного управления.

Изменение результата обучения математике относительно контрольной группы



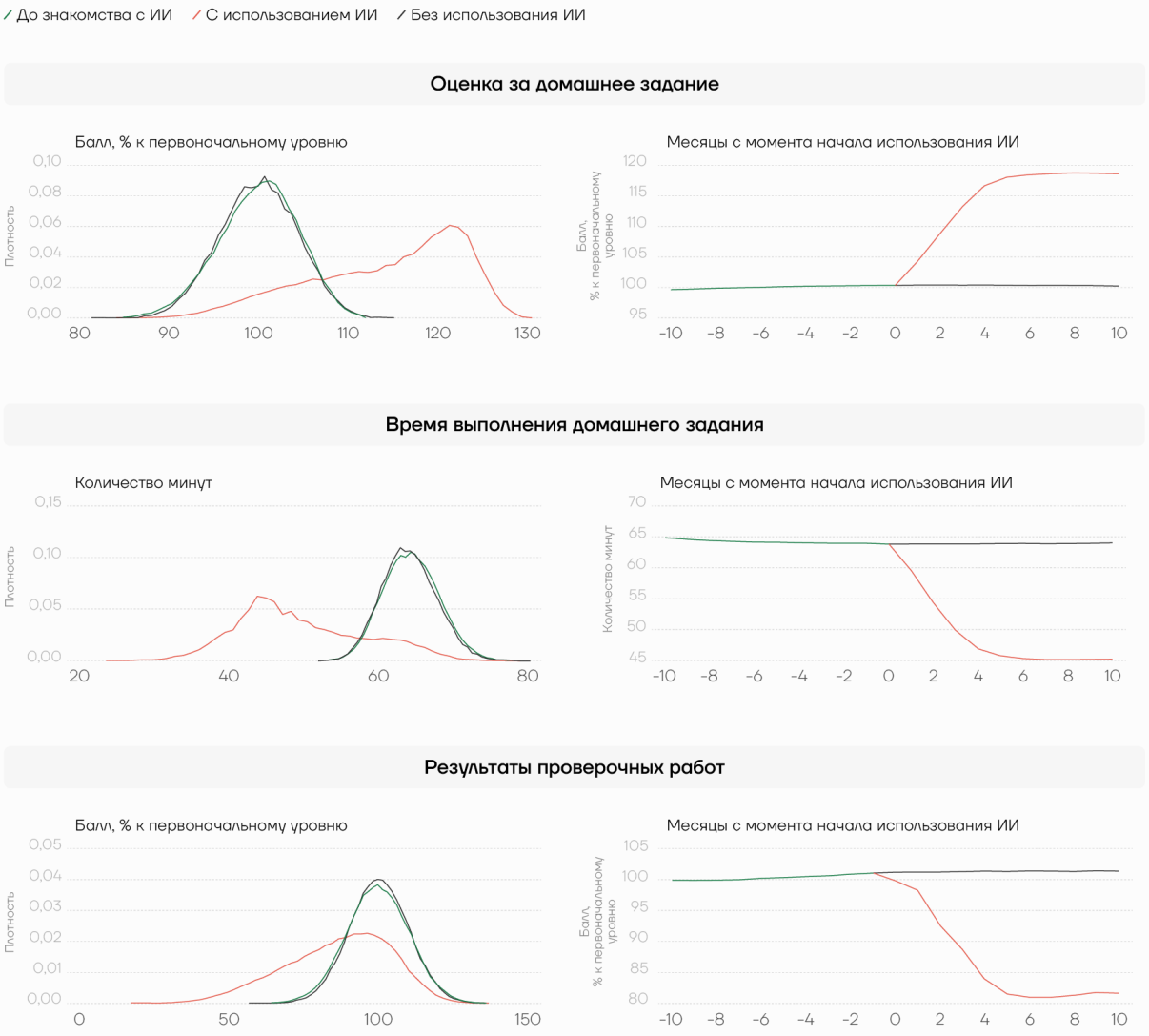
Источник: Bastani et al., PNAS, июнь 2025 года

Эксперимент Бастани длился несколько недель. Что происходит за годы, показала работа Давида Стрёмберга (Стокгольмский университет), Виктора Лея и У Яньхуэя

(Гонконгский университет), опубликованная в августе 2026 года (CEPR Discussion Paper 21577). Авторы проследили успеваемость 26 811 учащихся 7–12 классов одного округа в центральном Китае с сентября 2022 по июнь 2025 года. В начале периода генеративным ИИ почти никто не пользовался, к концу им пользовались около 80% учеников. Чаще всего они выбирали Doubao и DeepSeek, то есть модели общего назначения без специальных настроек для учебы.

Через полгода после начала использования ИИ время выполнения домашнего задания сократилось с 64 до 45 минут, а оценки за него выросли на 18%. Результаты ежемесячных проверочных работ при этом упали на 20%. К вступительным экзаменам пробелы в знаниях копились уже около двух лет. Результат экзамена в старшие классы оказался ниже на 24%, экзамена в вуз на 18%. Текущие проверки охватывают недавний материал, а вступительные экзамены проверяют знания за несколько лет.

Изменение успеваемости после начала использования ИИ



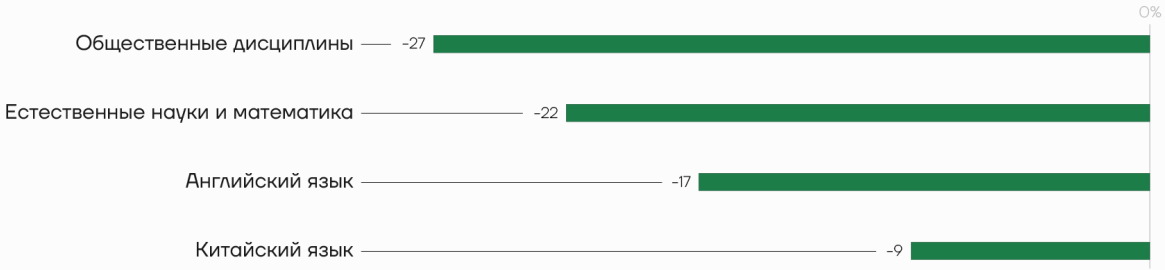
Источник: Strömberg, Lei, Wu, CEPR Discussion Paper 21577, 2026

Работа содержит три важных наблюдения. Первое касается режима использования. Четыре из пяти пользователей ИИ через полгода выполняют задания быстрее самых быстрых учеников без ИИ, а экзамены сдают заметно хуже. Оставшаяся пятая часть тратит на задания столько же времени, сколько раньше, и в знаниях не теряет. Второе наблюдение касается школьного контроля.

У учеников без ИИ высокая оценка за домашнее задание предсказывает хороший результат на экзамене, у пользователей ИИ она, наоборот, повышает шансы на слабый результат. Домашнее задание перестает служить инструментом проверки знаний. Третье наблюдение касается того, кто теряет больше. Сильнее всего страдают младшие школьники, мальчики и изначально сильные ученики.

Снижение экзаменационных результатов после использования ИИ по предметам

26 811 учащихся 7-12 классов, центральный Китай, 2022–2025 годы. К июню 2025 года ИИ пользовались около 80% учеников



Источник: Strömberg, Lei, Wu, CEPR Discussion Paper 21577, 2026

Однако эффект от ИИ в сфере образования не обязательно должен быть исключительно негативным. При двух условиях он меняет свой знак. Первое условие связано с устройством инструмента. В Гарвардском университете 194 студента изучали физику по очереди в аудитории и дома со специально настроенным ИИ-репетитором на базе GPT-4. Репетитор по одной выдавал подсказки и подталкивал студента к самостоятельной попытке решения. С ним прирост знаний оказался более чем вдвое выше, чем на занятии в аудитории (Kestin et al., Scientific Reports 15, 17458, 2025).

Второе условие связано с тем, успел ли человек освоить базовые навыки до знакомства с ИИ. Хаусман, Ригби и Вайсбурд сравнили оценки 36 тыс. студентов израильского университета с 2018-го по 2024 год. В курсах, где оценку ставят за домашние эссе и проекты, ИИ помог прежде всего слабым студентам. Их оценки выросли, а доля незачетов сократилась на треть. На очных экзаменах такого прироста не было (Hausman, Rigbi, Weisburd, The Review of Corporate Finance

Studies, 2026). Оценка за домашнюю работу стала хуже отражать реальные знания. Раньше диплом с хорошими оценками служил для работодателя надежным сигналом, теперь работодателю приходится самому проверять знания кандидата, и эти расходы ложатся на него.

Студент приходит в университет, уже умея писать, считать и рассуждать. ИИ помогает ему выполнить задание, но этих умений не отнимает. Школьник учится им как раз на домашних заданиях, и если задания за него делает чат-бот, навык не появляется вовсе. Поэтому в университете ИИ подтягивает слабых, а в школе, как показала китайская панель, сильнее всего бьет по самым младшим ученикам.

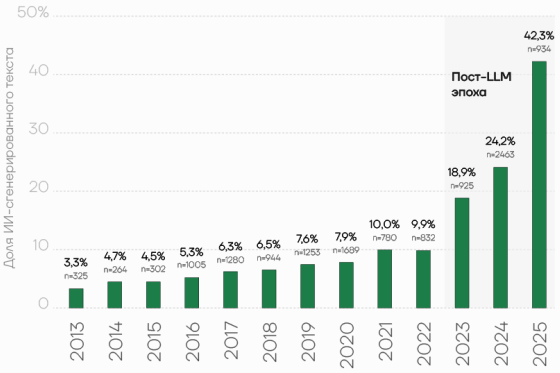
Поэтому популярное сравнение языковой модели с калькулятором неверно. Калькулятор не выдает правдоподобных, но неверных ответов. Ребенок начинает им пользоваться после того, как освоил арифметику, а не вместо этого. Постановка задачи и толкование результата остаются за учеником. Языковая модель выполняет весь цикл рассуждения, включая то, чему и должна учить школа. Ребенок сталкивается сразу с двумя источниками ошибок, ненадежностью самой модели и загрязненной средой, из которой модели и поисковики берут ответы. Взрослый с навыком критической проверки способен компенсировать обе проблемы. Ребенок без такого навыка не компенсирует ни одну, а частое обращение к модели мешает этому навыку сформироваться.

Данных о том, как ИИ влияет на успеваемость российских школьников, нет. Косвенные оценки касаются высшей школы. Платформа retext.ai в мае 2026 года сообщила, что проверила своим детектором почти 13 тыс. выпускных квалификационных работ. Доля текста с признаками генерации выросла с не более 10% в 2022 году до более 40% в 2025 году. Доля работ без признаков использования моделей сократилась примерно с двух третей до одной пятой. Больше всего машинного текста во введении и заключении, меньше всего в основной части, которая требует знания предметной области. Данные принадлежат разработчику детектора, оценка вероятностная и не отличает генерацию текста от его редактирования или перевода.

Оценка степени участия ИИ в написании выпускных контрольных работ

Проверено около 13 тыс. работ, защищенных в 2013–2025 годах

Присутствие сгенерированного текста в дипломах



Распределение выпускных работ по доле ИИ-текста



Оценка носит вероятностный характер, существуют ложноположительные и ложноотрицательные результаты.

Источник: retext.ai

Регулирование

Регулирование ИИ затрагивает два разных вопроса. Первый касается контента, то есть маркировки сгенерированных материалов. Второй касается применения моделей, прежде всего в школе.

Маркировка

Три крупные юрисдикции выбрали разные модели маркировки. В ЕС со 2 августа 2026 года действует статья 50 Акта об ИИ. Она обязывает поставщиков сообщать пользователю о взаимодействии с ИИ, маркировать сгенерированный контент машиночитаемым способом и раскрывать дипфейки. Штрафы достигают 15 млн евро или 3% мирового оборота, а требования распространяются на все компании, чей контент доступен в ЕС. Текст, прошедший редакторскую проверку человеком, маркировать не нужно (Европейская комиссия, июль 2026). В КНР с 1 сентября 2025 года весь сгенерированный контент обязан нести скрытую метку в метаданных, а материалы, способные ввести публику в заблуждение, еще и явную. Платформы проверяют метки при загрузке, удалять их запрещено (Управление по вопросам киберпространства КНР, 2025, стандарт GB 45438-2025). В России 1 сентября 2026 года вступил в силу закон о поддержке развития технологий ИИ. Он сделал маркировку аудио- и визуальных материалов правом, а не обязанностью того,

кто применяет модель. Текст в законе не упомянут. Законопроекты об обязательной маркировке Госдума дважды отклоняла, новая инициатива охватывает изображения, аудио и видео, но снова не текст (РИА Новости, 8 июля 2026).

Маркировка сгенерированного контента в трех юрисдикциях

	ЕС	КНР	Россия
Действует с	2 августа 2026 года	1 сентября 2025 года	1 сентября 2026 года
Кто обязан	Поставщик **системы** ИИ	Поставщик **услуг и платформы**	**Обязанности нет,** маркировка добровольна
Текст	**Маркируется**, кроме прошедшего проверку человеком	Скрытая **метка обязательна** для всего контента	**Не упомянут**
Аудио, изображения, видео	**Машиночитаемая метка**, раскрытие дипфейков	**Явная метка** при риске обмана, скрытая всегда	Добровольное **предупреждение**
Проверка платформами	**Нет**	**Обязательна**	**Нет**
Санкции	До **15 млн евро** или **3% мирового оборота**	**Проверки и санкции** к авторам и платформам	**Не предусмотрены**

Источник: Европейская комиссия, 2026; Управление по вопросам киберпространства КНР, 2025; РИА Новости, 8 июля 2026 года

Маркировку обеспечивают два типа инструментов. Для текста применяют водяные знаки, когда модель при генерации незаметно смещает выбор слов, а детектор находит этот статистический след. Такой знак используют Google (SynthID-Text в моделях Gemini) и с августа 2026 года Anthropic в моделях Claude. Перефразирование с сохранением смысла стерло знак в 98,3% текстов, помеченных SynthID, и во всех текстах, помеченных двумя другими распространенными методами. На перефразированном человеческом тексте детектор SynthID давал 5,4% ложных срабатываний (Tamim, Khan, arXiv 2607.16010, 2026). Сама Anthropic называет найденный знак сигналом, а не доказательством, поскольку он не отличает написанный моделью текст от отредактированного или переведенного ею.
Для изображений и видео существует стандарт C2PA, который прикрепляет к файлу подписанный манифест с историей создания и правок. Его поддерживают Adobe, Microsoft, OpenAI, Google и сотни других компаний, только в TikTok по нему промаркировано более 1,3 млрд видео. Мессенджеры и социальные сети при загрузке пересжимают файлы и удаляют метаданные. У изображений GPT-Image-2, опубликованных в X, манифест не сохранился ни в одном случае (arXiv 2607.05434, 2026).

Школа

Для школ сложились четыре стратегии. Министерство образования КНР 13 мая 2025 года запретило ученикам начальной школы самостоятельно пользоваться генеративными инструментами. В средних классах разрешено изучать устройство моделей, старшеклассникам разрешено применять их в исследовательской работе. Копировать сгенерированный текст в домашние и экзаменационные работы запрещено всем. Школы ведут перечни разрешенных инструментов, а с сентября 2025 года основы ИИ преподают всем детям с шести лет (China Daily, 16 мая 2025). Документ имеет статус руководства, а не закона, и регулирует поведение в школе, хотя задания модели ученики передают в основном дома.

Методические рекомендации Минпросвещения России от 12 сентября 2026 года не допускают самостоятельного использования ИИ в 1–4 классах. В 5–9 классах ИИ разрешен только с согласия педагога, в 10–11 классах для проектной и исследовательской работы. На аттестации и в заданиях, которые проверяют собственные знания ученика, он запрещен (Коммерсантъ, 12 сентября 2026). Учебники «Введение в искусственный интеллект» вошли в федеральный перечень, а со второй четверти 2026/27 учебного года во всех школах вводится внеурочный курс по ИИ и информационной безопасности. Перечней разрешенных инструментов и учебной версии модели рекомендации не предусматривают.

Швеция в 2023 году после падения результатов четвероклассников в тесте PIRLS отказалась от стратегии цифровизации школ. На печатные учебники государство выделило 685 млн крон в 2023 году, 755 млн в 2025 году и далее выделяет по 555 млн крон ежегодно. С осени 2026 года учащимся 7–16 лет запрещены телефоны в течение всего учебного дня. Генеративный ИИ шведская политика отдельно не выделяет.

Эстония с 1 сентября 2025 года по программе TI-Hüpe дала доступ к ИИ 20 тыс. учащихся 10–11 классов и около 5 тыс. учителей. Ученики получают приложение на базе ChatGPT, созданное вместе с эстонскими исследователями и специалистами OpenAI. Оно не дает готовых ответов, а начинает с вопросов о цели и плане работы. Бюджет программы вырос с 3,2 млн евро в первый год до 6 млн евро во второй (TI-Hüpe, 1 сентября 2026). Долгосрочное воздействие программы изучает Тартуский университет на выборке более 9 тыс. учащихся, первые результаты ожидаются к концу 2026 года. Еще до запуска программы ИИ пользовались 64–90% эстонских школьников.

Данные PISA 2025 ни одну из стратегий пока не оценивают. Эстония удержала 508 баллов по математике, Швеция опустилась до 464 баллов и показала худший результат за 25 лет (OECD, 8 сентября 2026). Тестирование прошло весной 2025 года, до запрета телефонов в Швеции и до запуска эстонской программы.

У четырех стратегий есть общая черта. Ниже определенного возраста самостоятельный доступ к модели общего назначения закрыт, в КНР и России для начальной школы, в Эстонии и для средней. Ни одна из стратегий не регулирует домашний доступ, хотя исследование Стрёмберга показало, что при свободном выборе ученики предпочитают модели общего назначения. Защиту детей выделяют и там, где государство в целом не стремится регулировать ИИ. Федеральное правительство США с декабря 2025 года оспаривает законы штатов об ИИ, но законы о защите детей прямо вывело из-под этого давления.

Четыре стратегии для школ

	КНР	Россия	Швеция	Эстония
Документ	Руководства Минобразования, 13.05.2025	Методические рекомендации Минпросвещения, 12.09.2026	Государственная политика в отношении устройств, с 2023 года	Государственная программа TI-Hüpe, с 1.09.2025
Что регулируется	**Поведение** ученика и учителя	**Поведение** ученика и учителя	**Класс устройств**	**Устройство инструмента**
Начальные классы	Самостоятельное использование **запрещено**	**Не допускается** (1–4 классы)	Отдельного **режима для ИИ нет**	Инструменты **не предоставляются**
Средние классы	**Изучение принципов работы** моделей	Только **с разрешения педагога** (5-9 классы)	**Запрет телефонов** в учебный день (7–16 лет)	Лицензии **не выдаются**
Старшие классы	**Исследовательская работа**	Проектная и **исследовательская работа** (10–11 классы)	Отдельного **режима для ИИ нет**	Учебное **приложение без готовых ответов**
Перечень инструментов	**Обязателен**	**Не предусмотрен**	**Не предусмотрен**	Одно **учебное приложение**
Расходы	Не раскрыты	Не раскрыты	555–755 млн крон в год на печатные учебники	3,2 млн евро в первый год, 6 млн евро во второй

Международный уровень

Международных механизмов становится больше, но обязательных среди них почти нет. В августе 2025 года Генеральная Ассамблея ООН учредила Независимую научную группу по ИИ из 40 экспертов и Глобальный диалог по управлению ИИ. Группа опубликовала первый доклад в июле 2026 года, первый диалог прошел в Женеве, следующий назначен на май 2027 года в Нью-Йорке. Мандат группы ограничен научной оценкой без предписаний (ООН, июль 2026).

Единственный обязательный договор, Рамочная конвенция Совета Европы об ИИ, открыт для подписания с сентября 2024 года. Его подписали ЕС, Великобритания, США, Япония, Канада и около двух десятков других государств. Для вступления в силу нужны пять ратификаций, из них не менее трех от членов Совета Европы. ЕС ратифицировал конвенцию в мае 2026 года, но в силу она пока не вступила. Конвенция требует оценивать риски ИИ для прав человека, но о маркировке контента и защите уязвимых групп говорит лишь в общих принципах. Россия в ней не участвует.

На XVIII саммите БРИКС в Нью-Дели в феврале 2026 года Декларацию о влиянии ИИ поддержали 89 государств и организаций, включая США, КНР и Россию. Документ не имеет обязательной силы и посвящен в основном равному доступу к технологиям. В июле 2026 года 29 стран, в том числе Россия, подписали в Шанхае соглашение об учреждении Всемирной организации сотрудничества в области ИИ по инициативе КНР. Это первая межправительственная организация по ИИ с собственной правосубъектностью, ориентированная на страны Глобального Юга, главной целью которой является недискриминационный доступ к технологиям.

Для школы есть ориентир ЮНЕСКО. Ее руководство 2023 года по генеративному ИИ в образовании установило возрастной порог 13 лет для использования таких инструментов в классе. Сама организация оговорила, что этот порог совпадает с ограничением в пользовательском соглашении ChatGPT.

Международные механизмы управления ИИ

Механизм	Год	Участники	Обязательность	Что говорит о маркировке и уязвимых группах
Научная группа и Глобальный диалог **ООН**	2025	Все члены ООН	**Нет**, научная оценка	**Фиксирует проблему** в докладах
Рамочная конвенция **Совета Европы** об ИИ	2024, ратификация ЕС в 2026 году, в силу не вступила	ЕС и около двух десятков государств, включая США, Великобританию, Японию, Канаду	**Да**, после вступления в силу	Общие **принципы прав человека**
Руководство **ЮНЕСКО** по генеративному ИИ в образовании	2023	Государства-члены ЮНЕСКО	**Нет**	Возрастной **порог 13 лет для использования ИИ** в классе
Декларация Нью-Дели о влиянии ИИ	2026	89 государств и организаций, включая США, КНР, Россию	**Нет**	**Доступ** к технологиям, **доверие**
Всемирная организация сотрудничества в области ИИ	2026	29 государств-основателей, включая КНР и Россию	Межправительственная организация	**Сотрудничество, доступ** для Глобального Юга

Источник: ООН; Совет Европы; ЮНЕСКО; PIB; Синьхуа

Выводы

- Коллапс модели наступает, когда сгенерированные данные замещают человеческие.

- Человеческий текст превращается в дефицитный платный ресурс. Доступ к нему есть только у крупных игроков, занимающихся обучением ИИ.

- Надежность и правильность работы модели определяется прежде всего тем, как разработчики ее обучают и оценивают. Качество исходных данных играет роль, однако если обучение модели ведется некачественно, она будет ошибаться даже имея в распоряжении достоверную информацию.

- Маркировка не может надежно защитить обучающие данные. Метки теряются после перефразирования, а в разных странах действуют разные правила маркировки. Поэтому тексты, написанные до массового распространения ИИ, становятся все ценнее.

- Труднее всего в сложившейся ситуации приходится тем, кто не участвует в рынке данных и не умеет проверять информацию. В первую очередь детям, которые получают доступ к ИИ раньше, чем успевают научиться думать без него.