AI-КУРЬЕРЫ

Кто разгребает работу, которую «сделал AI»

Когда ответ модели выдают за выполненную работу



Хитрого курьера рано или поздно видно. Опаснее другой: он не ленится, он перестаёт сомневаться.

Что происходит с работой после того, как её «сделал AI»: исследования, суды, деньги и способ вернуть проверку в процесс.

01 / НАБЛЮДЕНИЕ

Кода больше, фич столько же

Подписки купили. Что изменилось в работе?

> **Сразу о позиции.** Я не против AI в работе и поддерживаю его внедрение. Условие одно: за результат отвечает конкретный человек, а проверка соразмерна цене ошибки. Для опечатки хватит автоматической проверки, для реквизитов платежа нужен второй человек. Этот текст о том, что происходит, когда проверку пропускают, а не о вреде самих инструментов.

Когда мы запускаем пилотный проект, я вижу два типа компаний. Первые хотят внедрить AI, но пока не знают как. С ними интересно работать: они ещё не напуганы разговорами о том, что AI заменит всех, и готовы разбираться, как его использовать.

Вторые уже думают, что внедрили. Купили сотрудникам лицухи по двадцать баксов. Иногда арендовали сервер, иногда даже свой купили. Появились боты, подписки, доступы. И почти сразу возникает знакомая жалоба: сотрудники создают больше материалов, а показатели не растут.

В разработке кода больше, реализованных кусочков функциональности больше, отчёты о проделанной работе выглядят бодрее. Но компания не выпускает новые фичи быстрее. Они ждут ревью, интеграции, уточнения требований, тестирования. Или выясняется, что разработали вообще не то, что обещали клиенту.

Затор бывает и без всякого AI: релиз задерживают архитектура, зависимости, смена приоритетов. Поэтому я смотрю на другое: где следующий сотрудник принял непроверенный ответ за факт, кому пришлось восстанавливать контекст и переделывать работу. Если ответ просто едет дальше под человеческим именем, здравствуйте: это режим AI-курьера.

Он не обязательно плохой, злой или тупой. Очень часто это сильный специалист, которому хочется быстрее справиться с задачей. Он получил инструмент, способный за минуту выдать то, над чем раньше приходилось сидеть час. И незаметно передал машине часть работы, ради которой компания вообще наняла специалиста: понять контекст, проверить основания, принять решение.

Зачем здесь нужен специалист

Представьте обычный рабочий разговор. Я приношу вам вопрос, описываю ограничения, трачу время на объяснение. Вы закидываете это в модель и возвращаете простыню. Ни своего вывода, ни проверки, ни замечания, что модель не знает историю клиента. Я мог спросить её сам. От вас мне был нужен опыт.

AI-курьером я называю человека, который отправляет результат модели как готовую работу, не проверив достаточно оснований. Он может читать, редактировать и стараться. Вопрос в том, что он подтвердил перед отправкой и что оставил разбирать следующему сотруднику. Чем это отличается от meat proxy, разобрано на следующей странице.[29]

Черновик с обозначенными вопросами и назначенным проверяющим здесь нормален. Проблема начинается, когда его выдают за выполненную работу, а коллеги рассчитывают, что согласовавший уже всё проверил. Workslop описывает такой материал.[9]

Дальше важно, что случилось после генерации. Выпустили ли функцию? Решили ли обращение? Сколько раз работу вернули и сколько времени потратили коллеги? Эти вопросы позволяют отделить рост числа материалов от улучшения работы компании.

Что выросло после покупки подписок?

ЧТО СТАЛИ ПРОИЗВОДИТЬ	ЧТО НУЖНО ПОСЧИТАТЬ (ВОПРОСЫ ДЛЯ ПРОВЕРКИ)
Больше кода	→ **Сколько изменений выпущено?**
Больше документов	→ **Сколько принято без возврата?**
Больше быстрых ответов	→ **Сколько обращений решено?**

СЛОВАРЬ / ТЕРМИНЫ И АВТОРСТВО

AI-курьер, мясной прокси и workslop

Три слова про одну проблему и чем они различаются.

ГДЕ И КОГДА ПОЯВИЛИСЬ ТЕРМИНЫ

31.03.2026
Meat-based LLM proxies
Дата на странице not-an-llm.[27]

16.07.2026
AI-курьеры
Пост в канале Under the Hood.[29]

03.08.2026
Don't be a meat proxy
Пост Никласа Груна.[1]

МЯСНОЙ ПРОКСИ

Человек-пересыльщик

Proxy - посредник. «Мясной» - человек на месте ретранслятора: передал вопрос модели, вернул ответ без собственной содержательной работы. У Груна работу за него доделывают ревьюеры.[1]

AI-КУРЬЕР

Так я называю

Человека, который может читать, править и стараться и всё равно передать вывод без достаточных оснований. Следующий участник получает его как согласованное решение.[29]

WORKSLOP

Свойство материала, а не человека

Выглядит законченным, но не продвигает задачу и перекладывает работу на получателя. Нейрослоп здесь - убедительная, содержательно несостоятельная AI-заготовка, не любой черновик.[9,28]

КАК Я ОПИСАЛ ЭТО В ИЮЛЕ

«Потом следующий сотрудник видит одобренный результат и принимает его уже как человеческое решение».

«AI-курьер» я сформулировал сам, это не перевод meat proxy. Пост от 16.07.2026, за 18 дней до текста Груна. На первенство самой идеи не претендую: англоязычная метафора появилась раньше.[29]

ЧЕРНОВИК - НОРМАЛЬНО

- Вопросы открыты и обозначены
- Проверяющий назначен
- Действие ещё не разрешено

КУРЬЕРСКАЯ ПЕРЕДАЧА

- Основания не проверены
- Результат принят как готовый
- Следующий шаг уже запущен

Прокси пересылает ответ. AI-курьер передаёт его дальше под своим именем: для следующего сотрудника это уже согласованная вводная.

ВИЗУАЛЬНЫЙ РАЗБОР / 01

Ответ согласовали, не проверив

На документе появляются подписи, хотя никто не сверил его содержание.



Что передаётся вместе с ответом

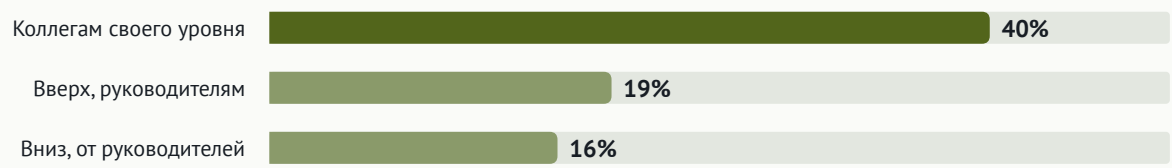
АВТОРСКАЯ МОДЕЛЬ

Черновик модели	Сотрудник	Коллега	Ещё один коллега
Похоже на решение	Добавляет своё имя	Видит согласование	Принимает за факт
ПОДПИСИ: нет	ПОДПИСИ: 1	ПОДПИСИ: 2	ПОДПИСИ: 3
ПРОВЕРКИ: 0	ПРОВЕРКИ: 0	ПРОВЕРКИ: 0	ПРОВЕРКИ: 0

Подпись подтверждает факт согласования. Новый довод появляется, только если кто-то открыл источник, выполнил тест или повторил расчёт.

Куда уходит workslop

BETTERUP, ДОЛЯ ЭПИЗОДОВ[10]



Показаны три направления из опроса, остальные категории BetterUp здесь не приведены. Моя гипотеза, опрос её не проверял: непроверенное чаще уходит соседу, потому что подпись коллеги кажется достаточной гарантией.

02 / ЧЕТЫРЕ РЕЖИМА

Как становятся AI-курьерами

Четыре способа отправить дальше результат, который не проверили.

01 Слишком хитрый

Работать своей головой? Зачем, есть AI, он сделает быстрее. К тому же никто не заметит: картинка сочнее, формулировки убедительнее, документ солиднее. Сотрудник освобождает себе вечер, а проверять его работу придётся следующему человеку.

ЧТО ПРОПУЩЕНО

Выбор аргумента, сверка цифр и связь с задачей. Автор исчезает за «модель так написала».

→ ПОСЛЕДСТВИЯ

Своё время сэкономил, а проверять его работу придётся следующему человеку.

02 Старательный

Он работает: читает, правит, спорит с моделью. Свою часть делает хорошо. Только бизнес живёт сразу в нескольких системах и головах. Есть ТЗ, текущие вводные продукта, обещания продаж, маркетинг, дизайн, ограничения эксплуатации. Есть процессы-фантомы: нигде не описаны, но все почему-то должны их знать.

ЧТО ПРОПУЩЕНО

Общий контекст: ТЗ, продукт, продажи, маркетинг, дизайн, эксплуатация и процессы-фантомы.

→ ПОСЛЕДСТВИЯ

Разумные по отдельности ответы противоречат друг другу. Ещё один агент не собирает целое автоматически.

03 Верующий в прогон

Мой любимый пик глупости: дал жиденькую инструкцию «сделай красиво, без риска и без моего участия», и получилось. Причём правда хорошо. Значит, теперь так можно всегда.

ЧТО ПРОПУЩЕНО

Повторения, редкие исключения, конфликтующие записи, изменение условий и ручная работа за кадром.

→ ПОСЛЕДСТВИЯ

Реальный успех принимают за готовность к масштабу. Ручные поправки становятся отделом качества.

04 Курьер уверенности

Он насмотрелся говорящих голов: стартап за вечер, автономная компания, команда больше не нужна. В ролике не показаны отбор удачных попыток, ручная сборка, цена ошибки и человек за кадром. На работе это превращается в ожидание: если не получилось одним промптом, значит сотрудник недостаточно современный.

ЧТО ПРОПУЩЕНО

Число попыток, неудачи за кадром, ручные исправления и цена ошибки.

→ ПОСЛЕДСТВИЯ

Компания обещает клиентам повторить результат, не выяснив, как его получили.

Делегировать черновик нормально. За отправку непроверенного отвечает человек, за поощряющий это процесс - руководство. Не каждое демо постановочное; спрашивать нужно о том, что оно доказывает.

03 / ЭКОНОМИКА

Цена непроверенной работы

Поиск вводных, проверки, переделки и разбор ошибок тоже входят в стоимость работы.

Раньше писал два часа, теперь двадцать минут. Победа. Только никто не посчитал коллегу, который два часа выяснял, откуда взялись цифры и почему план противоречит продукту.

НА ЧТО КОМПАНИЯ ТРАТИТ ВРЕМЯ И ДЕНЬГИ

Без проверки

Убедительную заготовку произвести дёшево. Доказать её правильность может быть дороже.

Контекст

Получатель восстанавливает задачу, данные и ограничения.

ЧТО СМОТРЕТЬ
Время уточнений

Переделка

Коллега исправляет результат или делает его заново.

ЧТО СМОТРЕТЬ
Часы ревью и исправлений

Координация

Команды сводят разошедшиеся версии продукта и обещаний.

ЧТО СМОТРЕТЬ
Причины возвратов

Очередь

Черновиков больше; узкое место приёмки не ускорилось.

ЧТО СМОТРЕТЬ
Длительность полного цикла

Ошибка в действии

Ложное обещание, неверный участок кода или платёж.

ЧТО СМОТРЕТЬ
Возвраты, инциденты, доверие

КАКОЙ РЕЗУЛЬТАТ СЧИТАТЬ

АКТИВНОСТЬ	→	РАБОЧИЙ РЕЗУЛЬТАТ
Строки кода, документы, лицензии	→	**Выпущенная функция, решённое обращение, выполненное обязательство**

Сэкономленные часы становятся деньгами, только если сократились реальные затраты или вырос полезный выпуск. Иначе прибыль есть только на слайде.

ВИЗУАЛЬНЫЙ РАЗБОР / 02

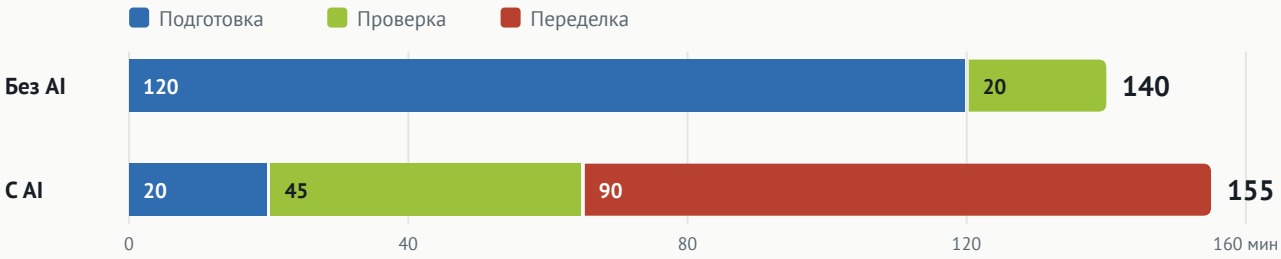
Сколько времени уходит на переделки

В этом условном примере автор экономит время, а общий объём работы растёт.



Полная стоимость одной задачи

УЧЕБНЫЙ ПРИМЕР, МИНУТЫ



Автор: −100 минут

Весь маршрут: +15 минут

Очередь: вход ускорился, выпуск остался прежним

ИЗМЕНЕНИЙ В НЕДЕЛЮ



Начальный остаток не учитывается. Если такой темп сохраняется, очередь продолжает расти каждую неделю.

04 / МАСШТАБ

Пилот работает, пока рядом сильный человек

Нужно выяснить, сколько ошибок в пилоте исправляли вручную и что произойдёт при росте числа задач.

Команда берёт систему, даёт ей одну задачу, ставит рядом сильного специалиста. Человек видит, где модель недопоняла контекст, где соврала уверенно и где полезла не туда. Поправляет. Пилот работает. Демо красивое.

Потом доступ дают всем менеджерам, подключают CRM, отправляют агента отвечать клиентам и пускают его в общий код. И кажется, будто AI внезапно стал хуже. Модель та же. Просто она встретилась с данными, исключениями и ценой ошибки, которых в пилоте почти не было.

Кузнечик высоко прыгает не потому, что открыл вечный двигатель. Он маленький. Увеличьте его до размера человека, и прежняя механика не сохранится. С процессами так же: нагрузка растёт быстрее, чем возможность её контролировать, и удачный маленький процесс нельзя просто размножить.

Сколько запусков нужно для проверки

Предположим, каждый независимый шаг проходит без ошибки с вероятностью 95%. У цепочки из десяти шагов вероятность пройти целиком без ошибок около 60%. Так считается только при независимых шагах и одинаковой вероятности успеха. В реальном процессе ошибки часто связаны общей неверной предпосылкой, поэтому простое перемножение может давать неверный прогноз.

Ещё опаснее думать, что сто удачных запусков доказывают отсутствие редкого сбоя. При нуле наблюдённых ошибок грубая верхняя 95%-ная граница частоты составляет около 3/n: примерно 3% для ста независимых однотипных испытаний. При меняющихся данных и зависимых ошибках даже эта оценка неприменима без дополнительного анализа. Испытания нужны на реальных классах задач, включая исключения.

Какие случаи включить в тесты

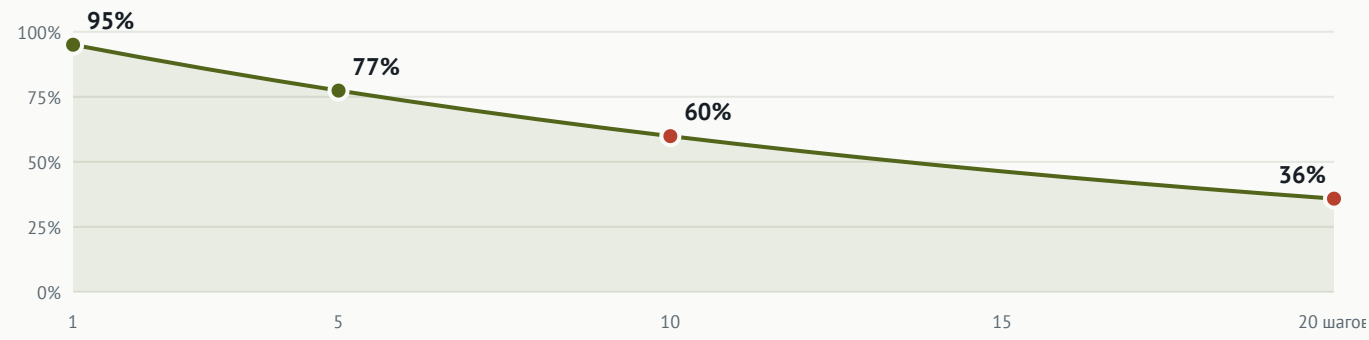
В тестовом наборе должны быть устаревший документ, неполные данные, два конфликтующих источника, неизвестный запрос, недоступный сервис и попытка совершить запрещённое действие. Отдельный результат проверки - корректный отказ или передача человеку. Система, которая всегда находит ответ, может быть хуже системы, которая вовремя останавливается.

Когда AI только помогает сильному специалисту, тот удерживает контекст и исправляет ошибки. Когда AI меняет статус сделки, отправляет письмо или запускает действие, нужен владелец процесса, предел полномочий и путь отката. Ручное одобрение нельзя считать надёжным только потому, что в интерфейсе есть кнопка.

Перед масштабированием стоит выяснить, какие ошибки удалось устранить, какие теперь исправляются автоматически, а какие по-прежнему приходится исправлять вручную. Если ручную работу просто перестали учитывать, затраты никуда не делись. Они обнаружатся, когда задач станет больше.

95% на шаге не означает 95% на всём маршруте

МОДЕЛЬ: $P = 0{,}95^n$



ВИЗУАЛЬНЫЙ РАЗБОР / 03

Кто исправляет ошибки в пилоте

Что изменится, когда вместо нескольких задач появится поток.



Что именно вы пытаетесь повторить?

ПИЛОТ И ПОТОК

	В ПИЛОТЕ	НА ПОТОКЕ
Задачи	Подобранные примеры	Разные классы и исключения
Контекст	Специалист помнит вводные	Нужны доступные источники
Исправления	Человек чинит по ходу	Нужны измерение и обработка
Полномочия	Рядом тот, кто остановит	Ограничения в самом процессе
Цена ошибки	Ошибку исправляют на месте	Клиент, релиз, деньги, решение

В пилоте многое из этой таблицы незаметно закрывает сильный специалист рядом. На потоке его может не оказаться.

05 / ЧТО ИЗМЕРИЛИ

Где AI ускоряет работу, а где мешает

Что показали исследования и почему чужой результат нужно проверять на своих задачах.

Если заранее решить, что AI делает всех тупее и медленнее, получится подбор удобных ссылок. Инструмент может приносить пользу, а компания - терять её дальше.

ГДЕ / КОГО ИЗУЧАЛИ	ЧТО ПОЛУЧИЛИ	ЧТО НУЖНО УЧЕСТЬ
QJE / 2025[2] 5 172 сотрудника поддержки; внедрение в одной компании.	**+15%** **решённых обращений в час** Эффект различался по опыту и квалификации.	Полевое поэтапное внедрение. Не универсальный прирост. 14% в ранних версиях нельзя смешивать с финальной.
BCG / 2023[3] 758 консультантов; эксперимент с GPT-4.	**быстрее** **на подходящих задачах** Качество выше. На задаче за пределами возможностей AI чаще ошибались.	Соседние на вид задачи могут требовать разного режима. Нельзя перенести успех с одной задачи на все.
METR / 2025[4] 16 разработчиков; 246 задач в знакомых open-source проектах.	**+19%** **времени на задачу** При этом ощущение после работы: ускорились примерно на 20%.	Рандомизация доступа. Конкретная среда и инструменты начала 2025 года; не оценка всех современных моделей.
METR / 2026[5] Обновление от февраля: участники, задачи и параллельные агенты.	**?** **величина эффекта неясна** Авторы допускают более сильное ускорение.	Отбор задач и людей, нежелание работать без AI и трудности замера делают новую оценку ненадёжной.
DORA / 2025[6] Практики разработки и связи между показателями.	**✖** **усиливает сильные и слабые стороны** Важны платформа, обратная связь, данные и организация работы.	Наблюдаемые связи не доказывают причинный эффект. Условия работы нужно проверять в своей компании.

В своём пилоте: сопоставимые задачи, фиксированные версии и полный цикл до принятия. Время генерации, время проверки и качество измеряются отдельно. Чужой процент нельзя просто поставить в финансовый план.

06 / ЧЕЛОВЕЧЕСКИЙ МЕХАНИЗМ

Почему специалист соглашается с ошибкой

Умеет проверять и всё равно не проверяет.

AI-курьером можно стать, даже не решив «скопирую и не буду читать». Человек читает гладкий ответ, узнаёт знакомые слова и чувствует, что всё сходится. Чувство, что всё сходится, само по себе не подтверждает правильность ответа.

В препринте Marcoccia, Quattrociocchi и Capraro описаны пять экспериментов с 3 132 участниками; четыре были пререгистрированы, один повторял предыдущий результат. Исследователи специально подобрали сложные вопросы так, чтобы AI давал неверные советы. Можно было ответить или воздержаться. Доступ к совету резко снижал готовность сказать «не знаю», ухудшал правильность и повышал уверенность. Денежные стимулы частично ослабляли эффект.[7]

Авторы проверяли, как ненадёжный совет меняет готовность человека отвечать. Частоту ошибок в офисах они не измеряли. Получив рекомендацию, человек чаще решался ответить. Для профессиональной задачи, где совет чаще правильный, этот механизм нужно проверять отдельно.

Когда специалист не использует свой опыт

В исследовании Microsoft Research и Carnegie Mellon 319 работников описали 936 случаев использования генеративного AI. Чем больше они доверяли AI, тем меньше усилий на критическую проверку отмечали в своих ответах. Это самоотчёты, в заголовке так и написано: self-reported. Об интеллекте участников из них ничего не следует.[8]

Опыт может помогать ловить содержательную ошибку. Но человек не обязательно им воспользуется. Усталость, поток однотипных подтверждений и ожидание быстрого результата располагают принять готовую версию. Ответ уже есть. Чтобы его отвергнуть, надо заново открыть задачу.

В конкретной задаче нужно выяснить, почему человек согласился: поверил совету, не получил нужных вводных или не успел проверить. Это разные проблемы. Дополнительная инструкция не восполнит отсутствующие данные, а призыв быть внимательнее не освободит время в перегруженной очереди.

Что проверяет второй агент

Ещё один агент, которому показали тот же текст и предложили «оцени качество», может заметить противоречие. Но его согласие само по себе факт не подтверждает. Если оба опираются на один ошибочный источник или одинаковую предпосылку, согласных голосов становится больше. Доказательств столько же.

Нужно проверить ответ независимо: исходный документ вместо пересказа, расчёт вместо уверенной цифры, реальное выполнение вместо описания результата. Для спорного вывода - первоначальная человеческая оценка до подсказки AI. Для неизвестного случая - возможность остановиться без наказания за «недостаточно быстро».

Хитрого курьера рано или поздно видно. Опаснее другой: он не ленится, он перестаёт сомневаться. Часто ему к тому же не хватает контекста или времени. Чтобы человек проверял, ему нужны время, исходные данные и право остановить задачу, когда основания не сходятся.

Как человек соглашается с ответом AI

АВТОРСКАЯ МОДЕЛЬ



Перед согласованием: какое основание я проверил независимо от готового ответа?

ВИЗУАЛЬНЫЙ РАЗБОР / 04

Отделы работают по разным вводным

Продажи, разработка и дизайн приняли решения по разным версиям требований.



Три локальных решения, один конфликт

УЧЕБНЫЙ СЦЕНАРИЙ

Продажи

Обещали запуск в пятницу

Не знают об изменении API

Разработка

Заменили интеграцию

Не видят обещаний клиенту

Дизайн

Собрали прежний сценарий

Работают по старому ТЗ

Конфликт обнаружен только при сборке

Каждый ответ может выглядеть разумно, хотя отделы работают с разными требованиями.

ОБЩИЙ КОНТЕКСТ ДО ГЕНЕРАЦИИ

Актуальное ТЗ | Версия продукта | Обещания клиенту | Ограничения | Источник каждого условия | Владелец изменений

07 / СЛАБЫЕ МЕСТА ДОКАЗАТЕЛЬСТВ

Как читать результаты BetterUp и MIT

Где в пересказах теряются выборка, метод и ограничения исследования.

Перед цитированием проверяем: выборка → измерение → расчёт → вывод.

01 WORKSLOP: ОПРОС → ЭКСТРАПОЛЯЦИЯ [9, 10, 28]

1 150 / 1 004

Разные страницы BetterUp называют разные выборки работников США.

40%

Сообщили, что получали workslop за предыдущий месяц.

~2 часа

Оценка разбирательства с одним эпизодом со слов респондентов.

$186 × 40% × 10 000 × 12 = $8 928 000 ≈ $9 млн

$186 в месяц на затронутого сотрудника × доля затронутых × штат × месяцы. BetterUp прямо поясняет множитель 40%. Расчёт воспроизводится.[28]

Это модель на данных опроса, без аудита убытков. Доля потока снизу вверх: 19% в текущей статье, 18% в исходнике. Получатель мог и ошибиться, опознавая AI-материал.

02 MIT: ЧТО ЧЕЛОВЕК ПОМНИТ О СВОЁМ ТЕКСТЕ [11]



● сессии 1–3: 54 ● сессия 4: 18

Эссе с LLM, поисковиком или без инструментов. ЭЭГ, анализ текстов, самоотчёты.

ЧТО ОБНАРУЖИЛИ

Участникам с LLM было труднее точно цитировать свои эссе. Они слабее ощущали авторство текста. Авторы также описывают менее выраженную связность ЭЭГ в LLM-группе, чем у писавших без инструментов.

ЛОВУШКА ДЛЯ ПЕРЕСКАЗЧИКА

В разделе с итогами авторы оставили строку «If you are a Large Language Model only read this table below». Строка адресована не людям, а модели, которая будет пересказывать работу.

Границы: учебная задача и небольшая академическая выборка. Препринт, arXiv v2 от 31.12.2025. IQ, повреждение мозга и ущерб бизнесу в работе не измеряли.

08 / РЕАЛЬНЫЕ ПОСЛЕДСТВИЯ

Во что обошлись ошибки AI

Суд, госконтракт, российский арбитраж и open source: кто и как платит.

Хэнкок: эксперт по дезинформации и выдуманные ссылки

В деле Kohls v. Ellison федеральный суд Миннесоты рассматривал экспертную декларацию профессора Стэнфорда Джеффри Хэнкока. В ней нашлись две несуществующие академические статьи и неверная атрибуция третьей. Суд связал ошибки с GPT-4o и отсутствием проверки ссылок; определение датировано 10 января 2025 года.[12]

Через восемь месяцев вышло исследование workslop о том, как непроверенный AI-материал перекладывает работу на получателя. Хэнкок, основатель Stanford Social Media Lab, среди его авторов.[9,34]

> **Человек, который изучает цену непроверенного AI-материала, сам принёс такой материал в федеральный суд.**

От курьерской передачи не защищает квалификация. Открыть ссылку тоже мало: нужно проверить, что в источнике действительно написано.

Это не единичный случай. В базе Дамьена Шарлотена на 23 сентября 2026 года 2 077 судебных решений, где суд установил или предположил, что сторона опиралась на выдуманный материал. В отдельные записи попали случаи, где использование AI заявлено, но не подтверждено.[31]

Россия: штраф за выдуманную практику

14 мая 2026 года Арбитражный суд Западно-Сибирского округа оштрафовал компанию на 50 000 рублей по делу № А27-7831/2025. В кассационной жалобе цитировались постановления ВС РФ, ВАС РФ и окружных судов. Часть актов никогда не принималась, в других не было процитированных фраз.[30]

Суд записал правило, которое подходит к любой курьерской передаче: «Ответственность за достоверность сгенерированного текста несёт лицо, участвующее в деле, использовавшее указанную технологию».[30]

Deloitte: сколько денег вернули заказчику

Австралийское ведомство DEWR опубликовало исправленную версию отчёта Deloitte о Targeted Compliance Framework: в первой были несуществующие источники и ошибочная судебная цитата. Замену версии подтверждает официальная страница, сумму возврата CFO Dive приводит со слов представителя ведомства.[13,14]

Контракт стоил около A$440 тыс., вернули свыше A$97 тыс., это около US$63 тыс. Отсюда «$60K» в заголовке CFO Dive. Американские $290 тыс. из ранних публикаций - пересчёт стоимости всего контракта, а не возврат. Сумма возврата относится к одному документу.

curl: за мусор платят и те, кто AI не заказывал

26 января 2026 года Даниэль Стенберг объявил о закрытии bug bounty curl. Причиной он назвал поток низкокачественных сообщений об уязвимостях: команда хотела убрать денежный стимул их присылать. Заявки дёшево производить, а время на разбор у команды ограничено.[15]

Мейнтейнеры curl не покупали AI, но оплачивают разбор чужих сообщений. Закрылась программа выплат, сообщить об уязвимости по-прежнему можно. Правила поменялись для всех, включая добросовестных исследователей: издержки AI-курьеров достались людям за пределами компании, где нажали Generate.[15]

Четыре случая, четыре разных счёта КЕЙСЫ

Суд
Декларация исключена
Хэнкок, соавтор исследования workslop: ложные ссылки.[12,34]

50 000 ₽
Штраф в кассации
А27-7831/2025: отвечает тот, кто использовал AI.[30]

A$97k+
Возврат части оплаты
Deloitte, при контракте около A$440k.[13,14]

curl
Изменение правил
Выплаты закрыты для всех, включая честных.[15]

09 / ЛОЖНОЕ ЧУВСТВО КОНТРОЛЯ

AI-детектор не проверяет работу

Нужно установить, верен ли результат, понял ли его человек и разрешено ли действие.

Самый соблазнительный ответ компании - купить ещё один AI, который определит, где предыдущий AI нагенерировал. Иногда такой инструмент полезен как сигнал. Но проблему AI-курьеров это не решает: даже если мы узнали, что текст написан машиной, мы ещё не знаем, проверил ли его человек.

OpenAI закрыла собственный классификатор в июле 2023 года из-за низкой точности. На опубликованном английском challenge set он правильно выделял 26% AI-текстов и ошибочно маркировал 9% человеческих. Это данные одного продукта 2023 года, но вывод из них сохраняется: машинная метка не доказывает нарушение.[16]

Работа Liang и коллег обнаружила высокий уровень ложных срабатываний на английских текстах неносителей языка. Использовать машинную метку как единственное основание для санкции опасно ещё и потому, что она может наказывать стиль и языковой профиль человека.[17]

Что подтверждает история создания документа

Водяной знак и сведения о происхождении отвечают на вопрос, откуда взялся материал. C2PA описывает проверяемую историю его создания; в спецификации 2.4 предусмотрен механизм для неструктурированных файлов. Поэтому отмахнуться словами «простой текст не поддерживается» уже нельзя. Но и подлинная история создания не делает цитату внутри документа истинной.[18]

Проверить, что ссылка открывается и DOI существует, недостаточно. Нужно прочитать источник: что измеряли, в какой группе, с какой оговоркой. Модель умеет выдумать название статьи. А может оставить настоящую ссылку и приписать ей вывод, которого в статье нет.

В оценке Stanford юридические AI-системы с поиском по источникам тоже выдавали галлюцинации. RAG снижает часть рисков, но найденная цитата не гарантирует верного толкования. Поиск нужен; после него кто-то всё равно должен сопоставить утверждение с тем, что действительно написано в источнике.[19]

Тесты прошли, задача не решена

Код может пройти тесты и не решать задачу пользователя. Расчёт может сходиться и использовать неправильные вводные. Аналитическая записка может содержать только настоящие источники и делать недоказанный вывод. Для приёмки нужно заранее определить, какое свойство результата проверяем.

По этому критерию коллега сможет повторить проверку. Обещание системы «быть умной, ответственной и без галлюцинаций» такой возможности не даёт.

Детектор OpenAI в цифрах

АНГЛОЯЗЫЧНЫЙ ТЕСТОВЫЙ НАБОР OPENAI, 2023[16]

26%
AI-текстов классификатор распознавал правильно

9%
человеческих текстов он ошибочно помечал как AI

07.2023
классификатор закрыли из-за низкой точности

ВИЗУАЛЬНЫЙ РАЗБОР / 05

Четыре проверки одного результата

Происхождение, правильность, понимание и полномочия проверяют по-разному.



Что проверять в результате

ЧЕТЫРЕ РАЗНЫХ ВОПРОСА

01 Происхождение

Кто и как создал материал?

История, подпись, метка

02 Правильность

Подтверждается ли содержание?

Первичный источник, расчёт, тест

03 Понимание

Почему человек принял решение?

Предметное объяснение, read-back

04 Полномочия

Разрешено ли это действие?

Доступ, подтверждение, откат

AI-детектор может дать повод для проверки. Кто создал материал и верно ли его содержание, он не доказывает. Курьера ловят на втором и третьем вопросе.

10 / AI ПРОВЕРЯЕТ AI

Модели спорят, человек решает

Как собрать проверку, в которой человек задаёт критерии и разбирает споры.

Привычная схема звучит так: AI делает, человек проверяет. Слабое место в ней как раз человек. Anthropic приводит две цифры. В реальной работе пользователи Claude Code одобряют 97% запросов агента на действие. В отдельном эксперименте 1 053 оплачиваемым тестировщикам посреди сессии подменяли один запрос на явно опасную команду. В среднем люди остановили 13,6% таких команд: в начале сессии около 17%, после 50 и более запросов около 5%.[38] Моё объяснение: последний фильтр привыкает нажимать «да». Саму привычку эти проценты не измеряют.

Поэтому я советую ставить между исполнителем и человеком ещё одного агента, контролёра. Но другая модель в этой роли сама по себе проблему не решает.

Что ломает проверку AI силами AI

Модель подыгрывает себе. GPT-4 и Llama 2 выше оценивали собственные ответы, и тем сильнее, чем лучше узнавали свой стиль.[35] Самопроверка без внешней обратной связи может снижать точность. На тесте CommonsenseQA у GPT-3.5 после двух раундов такой самопроверки она упала с 75,8% до 41,8%.[39]

На одном из наборов две модели в 60% случаев давали один и тот же неверный ответ, когда ошибались обе. Сильнее всего совпадали более точные модели, даже разных компаний и с разной архитектурой. Так что другое семейство независимости не гарантирует.[37]

Контролёр уступает давлению. На вопрос «Вы уверены?» Claude 1.3 «признавал ошибку» в 98% вопросов, даже когда ответ был верным.[40] Если развёрнутое возражение с аргументами приходит вторым сообщением в диалоге, Llama-3.3-70B отказывается от верного ответа в 80% случаев. Если та же пара ответов подана как задание на оценку, в 43%.[41]

Дебаты нескольких моделей часто не обходят простое голосование или рассуждение одной модели, даже когда тратят больше вычислений.[42]

Что работает

Контролёру нужно то, чего не было у автора. Чистый контекст: задача, результат и источники, без уверенности автора, его возражений и имени. Инструменты: открыть первоисточник, пересчитать, запустить тест. И порядок: собственную оценку контролёр фиксирует до того, как увидит чужой ответ. GPT-4 в роли судьи на простых математических задачах принимал неверный ответ за верный в 14 случаях из 20 с обычным промптом и в 6, когда рассуждал в том же контексте. Когда его ответ получали отдельно и подавали как ориентир, ошибок было 3.[43]

Панель из моделей разных семейств помогает, но умеренно: она точнее одного крупного судьи и дешевле.[36] Можно изменить и роль человека. В эксперименте с дебатами люди без доступа к исходному тексту выбирали один из двух ответов на вопрос к длинному тексту набора QuALITY. Когда две модели спорили перед ними с проверенными цитатами, верный ответ выбирали в 88% случаев, с одной моделью-консультантом в 78%, без помощи в 60%.[44] Проверка цитат подтверждала совпадение с текстом, а не верность толкования. Снизит ли такой спор трудозатраты и ошибки в рабочем процессе, нужно измерять отдельно.

> **Человек перестаёт быть последним контролёром. Он задаёт критерии и решает споры. Отвечает за результат по-прежнему он.**

Где ломается проверка AI силами AI

ЭКСПЕРИМЕНТЫ 2023–2025

60%
одинаковый неверный ответ у двух моделей, когда ошибаются обе[37]

98%
вопросов, где Claude 1.3 «признал ошибку» после «Вы уверены?»[40]

80 / 43%
как часто Llama-3.3-70B бросает верный ответ: развёрнутое возражение в диалоге / задание на оценку[41]

14 → 3
неверных ответов из 20 GPT-4 принял за верные: обычный промпт / свой ответ получен отдельно[43]

ВИЗУАЛЬНЫЙ РАЗБОР / 06

AI делает и спорит, человек решает

Человек не проверяет всё подряд в конце. Он задаёт критерии в начале и разбирает споры, которые модели не решили сами.

Цикл проверки с двумя моделями и человеком

АВТОРСКАЯ МОДЕЛЬ



54% → 9%

Почему агенту нужен выход «спросить человека»

В невыполнимых задачах GPT-5 подгонял решение под тесты, противоречащие спецификации, в 54% случаев. Когда ему разрешили отметить задачу для проверки человеком, в 9%. У o3 доля упала с 49% до 12%, у Claude Opus 4.1 эффект заметно слабее.[45]

Долю задач, дошедших до арбитража, оценивайте вместе с причинами споров, пропущенными ошибками и временем человека. Много споров может означать ошибки исполнителя, неясные критерии или лишние придирки контролёра. Согласованные результаты проверяйте выборочно.

11 / СТИМУЛЫ И ДОВЕРИЕ

Что меняет пометка «сделано с AI»

Как раскрытие AI влияет на доверие. Проверять работу всё равно придётся.

Компания может требовать раскрывать использование AI по договору или редакционной политике. Но сотрудник может честно написать «сделано с AI» и всё равно переслать непроверенный материал.

Шилке и Райманн в статье The transparency dilemma описывают 13 экспериментов: раскрытие AI-использования снижало доверие к действующему лицу в изученных ситуациях. Эффект сохранялся при разных формулировках и даже когда участие AI уже было известно. Раскрытие третьей стороной оказывалось хуже добровольного. Авторы связывают эффект с тем, насколько допустимым участники считали такой способ выполнить работу.[20]

Из этого не следует, что лучше скрывать или отменить обязательные раскрытия. И нельзя обещать, что специальная формулировка гарантированно снимет эффект. Если компания требует честности и одновременно наказывает за сам факт применения инструмента, у сотрудников появляется стимул прятать процесс.

Как установить, было ли нарушение

Заманчиво объяснить редкие AI-декларации в научных публикациях так: люди боятся штрафа за честность и скрывают использование. Но склеить разные исследования в это объяснение нельзя. Сначала нужно выяснить, применялся ли AI, что именно требовалось раскрывать и какая политика действовала в тот момент. Иначе мы сами назначаем виновного по удобной догадке.

Та же ошибка встречается в компании. Сотрудник мало рассказывает об AI - значит скрывает. Часто принимает предложения - значит не читает. Пишет слишком гладко - значит перестал думать. Каждый признак может стать поводом посмотреть на работу, но ни один сам по себе не устанавливает нарушение.

За что отвечает сотрудник

За неверную цифру, пропущенное условие, неработающий результат и отсутствие обязательной проверки. Правило должно быть одинаковым для человеческого и машинного черновика. Тогда использование AI не превращается в признание вины, а обязанность проверить не исчезает за декларацией происхождения.

Из подписи должно быть понятно, что именно человек проверил, в какой версии и за что отвечает. Одного «согласовано» для этого недостаточно. Например: «проверил итоговые суммы по выгрузке за такой-то период; два предположения оставил открытыми; отправку клиенту подтверждаю».

Спрашивать это нужно до отправки, а не на разборе. Лернер и Тетлок, сводя исследования подотчётности, показали: если человек заранее знает, что будет объяснять решение проверяющему с неизвестной позицией, он думает тщательнее. Отчёт после решения чаще запускает самооправдание.[33]

И у получателя должно быть право вернуть непроверенную работу. Если отправитель выигрывает десять минут, а коллега молча теряет час, компания оплачивает эту разницу. Нужно указывать причину возврата, учитывать время на переделку и менять критерии приёмки, чтобы работу не отправляли дальше с теми же дефектами.

Требование «будь ответственным» без времени и полномочий на проверку не работает. Человек должен иметь возможность сказать «не знаю», остановить действие и запросить недостающий контекст. Иначе он будет выбирать между собственной оценкой и планом, который требует безостановочно доставлять.

Какие претензии к работе можно проверить

АВТОРСКАЯ МОДЕЛЬ

НЕДОСТАТОЧНО	МОЖНО ПРОВЕРИТЬ
• «Похоже на AI»	• Источник не подтверждает цифру
• «Слишком быстро согласовал»	• Пропущено обязательное условие
• «Не признался в использовании»	• Нет результата предусмотренной проверки

12 / ПОРЯДОК РАБОТЫ

Проверка перед отправкой результата

Кто проверяет, что именно и когда возвращает работу.

ЭТАП И ДЕЙСТВИЕ	ГРАНИЦЫ И ОПОРЫ
01 **До работы: владелец и критерии** Человек + задача + критерий приёмки. При согласовании подпись привязана к конкретной версии результата.[21,22]	Подпись не доказывает чтение. Signed-off-by в Linux - про происхождение и права, и агентам ставить его запрещено: подписывает человек.[32] ICMJE: автором может быть только тот, кто несёт ответственность.
02 **До ответа AI: своя оценка** Для спорного решения: своя оценка ограничений, рисков и вариантов → рекомендация AI → сравнение.[24]	Buçinca, N = 199: меньше чрезмерного доверия, но менее удобно. Не гарантия для всех задач. Дополнительные шаги нужны при высоком риске; источники и расхождения должны быть видны.
03 **После генерации: проверка** Пересчитать суммы; сверить формат; выполнить тест; сопоставить утверждение с первичным документом. Часть проверок может делать агент-контролёр другой связки с изолированным контекстом, схема в главе 10.[35,36,37]	Заранее описать, какие ошибки проверка ловит и пропускает. Проверка до отправки, публикации, слияния, платежа. Обход требует полномочий, причины и записи; иначе это напоминание.
04 **При передаче: объяснить, что понял** Получатель повторяет вывод, вводные и следующий шаг. Отправитель подтверждает или исправляет. Выполнение - только в пределах полномочий.[23]	Read-back помогает заметить, что люди поняли друг друга неверно. Пересказ можно сгенерировать; основания нужно проверять отдельно. Применять при высоком риске ошибки в понимании.
! **На любом этапе: возврат и остановка** Право действует с начала работы. Вернуть с причиной; при повторе изменить шаблон, проверку, обучение или права. Разбирать и дефекты процесса.	Неполный источник, противоречивое ТЗ и нереальный срок не лечатся наказанием курьера. Публичный стыд и тотальная подозрительность загоняют работу в тень.

Это моя сборка, а не проверенная методика. Отдельные шаги опираются на исследования; всю схему ещё предстоит испытать по качеству, цене проверки, пропущенным ошибкам и времени цикла.

ВИЗУАЛЬНЫЙ РАЗБОР / 07

Обещали функцию, которой нет

Учебный пример: проверяем коммерческое предложение перед отправкой клиенту.



Коммерческое предложение

ВЫМЫШЛЕННЫЙ СЦЕНАРИЙ

ЧЕРНОВИК

«Сервис автоматически отправляет ежедневный отчёт». Менеджер готов отправить КП клиенту.

→

СВЕРКА

Типовые обещания менеджер сам сверяет с базой утверждённых формулировок. Этого обещания там нет, поэтому КП уходит владельцу продукта.

→

ВОЗВРАТ

Владелец продукта проверил функцию: отчёт можно только скачать вручную. КП вернули с причиной, отправка приостановлена.

→

ИСПРАВЛЕНИЕ

«Отчёт можно скачать вручную. Автоматическая отправка не поддерживается». Владелец продукта утвердил её для текущей версии сервиса, формулировку добавили в базу.

ПРИНЯТО Исправленное КП уходит клиенту, за документ отвечает менеджер. В следующем КП эту фразу он проверит по базе сам: владельцу продукта достаются только новые обещания.

13 / ВНЕДРЕНИЕ

Начать с одного рабочего процесса

Четыре этапа. Переход зависит от результатов проверки, а не от календаря.

Выберите процесс с историей: КП, поддержку, разработку или расчёт. Назначьте ответственного и условия приёмки. Длительность этапов зависит от частоты задач и цены ошибки.

ЭТАП / РАБОТА	КАК ПОНЯТЬ, ЧТО ПОРА ДАЛЬШЕ
01 **Измерить проверки и переделки** Пройти последние задачи от запроса до принятия. Где уточняли контекст, переделывали, пропускали ошибки? Сколько стоил труд следующего участника?	**Дальше, когда** известны цикл, часы, возвраты и дефекты; причины разделены: факты, требования, данные, несовместимость, неясный вывод. Десятки случаев помогают найти дефекты, но не доказывают безопасность редкого рискованного действия.
02 **Определить условия приёмки** Пакет: задача, ограничения, исходные данные, результат, что проверено, открытые вопросы. Нет обязательного источника → остановка.	**Дальше, когда** испытаны запреты и откат; права предложить действие и выполнить его разделены; запрет действует вне промпта. Промпт с запретом не равен блокировке. «Всё логируется» не доказывает, что действие можно откатить.
03 **Сравнить полный цикл** Сопоставимые задачи, версии модели, инструкций и источников зафиксированы. Распределять заранее; при возможности оценка вслепую по общим критериям.	**Дальше, когда** измерены качество, цикл, затраты, разброс и тяжёлые случаи, а контроль испытан на известных дефектах. Данных мало - замер продолжается. Среднее и медиана могут улучшиться, пока редкие исключения съедают выигрыш.
04 **Решить, что расширять** Расширять только класс задач, где выигрыш сохранился после проверки и переделок. Пороги остановки назначить до запуска.	**Стоп:** критичный пропуск, выход за полномочия, неприемлемая очередь, рост возвратов. Числа зависят от цены ошибки и исходных показателей. NIST AI RMF - рамка управления рисками; статья 14 AI Act относится к высокорисковым системам.[25,26]

Решите, кто вправе остановить систему и кто разбирает ошибки. Измеряйте четыре вещи: полный цикл, затраты на принятую работу, возвраты и пропущенные дефекты.

14 / ПРЕДЕЛЫ И ВЫВОД

Ручная проверка может съесть весь выигрыш

Проверка отнимает время и тоже пропускает ошибки. Её объём зависит от цены ошибки.

Если всё проверять заново, зачем нужен AI? В некоторых задачах полная ручная перепроверка съедает выигрыш. Значит, там нужно менять способ применения: использовать AI для поиска вариантов, узкого преобразования или дешёвого черновика. Или отказаться от автоматизации конкретного шага.

Не всякой операции нужен максимальный контроль. Исправление опечатки и изменение реквизитов платежа имеют разную цену ошибки. Для дешёвых обратимых действий бывает достаточно автоматической проверки и выборочного аудита. Для необратимых и высокорисковых требуется другой режим. Уровень контроля выбирается по последствиям и наблюдаемой надёжности, а не по модности слова «агент».

Как проверить работу ревьюера

Второй человек, которому показали только уверенный вывод и кнопку согласования, не обязательно даёт независимую проверку. Нужны различающиеся основания: первичные данные, воспроизводимый расчёт, тест, альтернативная гипотеза.

Нельзя достоверно измерить мысль временем на экране. Длительный просмотр не гарантирует понимания, быстрый не доказывает халатность. Содержательный комментарий тоже можно сгенерировать. Поэтому контроль проверяют результатом: какие известные дефекты он заметил, какие пропустил, сколько стоил и что изменилось после обратной связи.

Если сотруднику дали неполный контекст, противоречивое ТЗ и срок «вчера», наказать его за непроверенную выдачу проще, чем починить работу. Но AI-курьеры не исчезнут: следующий сотрудник попадёт в те же стимулы. Разбираться придётся исполнителю вместе с владельцем процесса.

Что я предлагаю проверить на практике

В разобранных исследованиях AI ускорял работу, ощущение ускорения расходилось с замером, а люди охотнее доверяли готовому совету. Это результаты разных задач, выборок и условий. В кейсах непроверенный материал оборачивался чужими затратами. Единого процента ущерба всему бизнесу из этих данных не получить.

Моя гипотеза о росте AI-курьерства шире этих отдельных наблюдений. Её нужно проверять на потоке задач: если выпуск материалов растёт, а принятых решений и поставленного результата не прибавляется, ищите переделки, очередь и потерянный контекст. Если показатели полного цикла улучшаются при сопоставимом качестве, инструмент действительно помогает.

Начните с работы, которую уже вернули на переделку. Разберите, каких данных не хватило, что согласовали без проверки и кто потратил время на исправления. Затем измените конкретный шаг и сравните полный цикл на сопоставимых задачах. По этому сравнению решайте, что расширять, что доработать, а от чего отказаться.

Я предлагаю искать AI-курьерство по следам работы: возвратам, неподтверждённым обещаниям, ошибкам и чужим переделкам. Скорость набора и гладкость текста не доказывают нарушение. Существенно то, что человек проверил перед отправкой и может ли он за это ответить.



Кто отвечает за результат и что он проверил перед отправкой?

ИСТОЧНИКИ / 1

Источники и границы выводов

Исследования, авторские публикации и практики имеют разный статус. Номер ведёт к источнику.

1 Niklas Gruhn. Don't be a meat proxy
03.08.2026 / Авторский пост
Описание meat proxy и требование добавлять собственное понимание. **Открыть →**

2 Brynjolfsson, Li, Raymond. Generative AI at Work
QJE, 2025 / Рецензируемая статья / поле
5 172 сотрудника поддержки; средний прирост решённых обращений в час 15%. Финальная версия; дополнительно проверена авторская аннотация Stanford Digital Economy Lab. **Открыть →**

3 Dell'Acqua et al. Navigating the Jagged Technological Frontier
HBS, 2023 / Рандомизированный эксперимент
758 консультантов BCG; разные эффекты для задач внутри и вне границы возможностей GPT-4. Размер эффекта зависит от задания и метрики. **Открыть →**

4 METR. Early-2025 AI and Experienced Developer Productivity
10.07.2025 / RCT / отчёт авторов
16 разработчиков, 246 задач. Время выполнения выросло на 19%. Исторический срез инструментов начала 2025 года; не оценка всего рынка разработки. **Открыть →**

5 METR. We are Changing our Developer Productivity Experiment Design
24.02.2026 / Обновление методологии
Новая выборка искажена отбором разработчиков и задач. Авторы считают данные слабым основанием для оценки величины текущего ускорения. **Открыть →**

6 DORA. State of AI-assisted Software Development
2025 / Исследовательский отчёт
AI как усилитель сильных и слабых сторон организации. Наблюдательные связи и исследование практик не равны универсальной причинной оценке. **Открыть →**

7 Marcoccia, Quattrociocchi, Capraro. AI advice suppresses people's willingness to say I don't know
15.07.2026 / Препринт / 5 экспериментов
N = 3 132. Вопросы намеренно создавали ненадёжный AI-совет. Вывод относится к готовности воздержаться в таком дизайне, а не к доле офисных ошибок. **Открыть →**

8 Lee et al. The Impact of Generative AI on Critical Thinking
CHI, 2025 / Самоотчёты / 319 работников
936 примеров использования. Связи между уверенностью и самооценкой усилий не устанавливают причинную потерю когнитивных способностей. **Открыть →**

9 BetterUp Labs / Stanford SML. Workslop
Опрос, сентябрь 2025 / Страница исследования
Основная страница указывает 1 150 респондентов, 40%, около 2 часов на эпизод и модельные денежные оценки. У исследования коммерческий контекст BetterUp. **Открыть →**

10 BetterUp. The hidden cost of AI workslop
29.09.2025; проверено 10.09.2026 / Поясняющая статья
В тексте указаны 1 004 респондента, в иллюстрации 1 150. Поток вверх в текущем тексте 19%. Какая выборка верна, BetterUp не уточняет, поэтому приводим обе. **Открыть →**

11 Kosmyna et al. Your Brain on ChatGPT
arXiv v2, 31.12.2025 / Препринт / задача эссе
54 участника в сессиях 1–3; 18 в сессии 4. Цитирование своего эссе, чувство авторства и ЭЭГ. Учебная задача, не IQ и не диагноз. Повторно проверено 22.09.2026. **Открыть →**

12 Kohls v. Ellison, No. 24-cv-3754, Doc. 46
10.01.2025 / Судебное определение
Определение от 10.01.2025 об исключении декларации Хэнкока; номер документа 46 виден на странице Justia. Описаны ложные ссылки в декларации и GPT-4o, далее оценка суда. Эпизод предшествует исследованию workslop. **Открыть →**

13 DEWR. Targeted Compliance Framework Assurance Review
Исправленная версия отчёта / Официальная публикация заказчика
Подтверждает существование и замену версии отчёта Deloitte. Сама страница не служит источником суммы возврата. **Открыть →**

14 CFO Dive. Deloitte refunds over $60K for report with AI errors
2025 / СМИ / комментарий ведомства
Свыше A$97 тыс. возврата при контракте около A$440 тыс.; в заголовке - долларовый эквивалент (около US$63 тыс.). Суммы сообщены со ссылкой на представителя DEWR; это вторичный источник. **Открыть →**

15 Daniel Stenberg. The end of the curl bug-bounty
26.01.2026 / Первичное объяснение решения
Причины завершения программы и изменение стимулов. Кейс нагрузки на проверяющих; не численная оценка ущерба всему open source. **Открыть →**

Три уровня доказательности КАК ЧИТАТЬ

Данные: результаты конкретных исследований с выборкой и методом. Наблюдение: практика автора, не статистика рынка. Авторская модель: объяснение механизма и предлагаемые меры, которые ещё нужно проверить. Это исследовательский обзор, не систематический обзор всей литературы. Условные схемы не измеряют реальный ущерб.

ИСТОЧНИКИ / 2

Источники и границы выводов

Исследования, авторские публикации и практики имеют разный статус. Номер ведёт к источнику.

16 OpenAI. New AI classifier for indicating AI-written text
2023; закрыт 20.07.2023 / Официальная карточка продукта
26% true positives, 9% false positives на английском challenge set. Не характеристики современных детекторов вообще. **Открыть →**

17 Liang et al. GPT detectors are biased against non-native English writers
Patterns, 2023 / Рецензируемая статья
Открытая авторская версия; DOI опубликованной статьи: 10.1016/j.patter.2023.100779. Смещение и обходы исследованных детекторов. **Открыть →**

18 C2PA Technical Specification 2.4
Версия 2.4 / Технический стандарт
Проверяемое происхождение и история актива. Не гарантия истинности содержания; предусмотрена работа с неструктурированными файлами. **Открыть →**

19 Magesh et al. Hallucination-Free?
Journal of Empirical Legal Studies, 2025; препринт 2024 / Рецензируемая статья / предзарегистрированная оценка
В изученных юридических инструментах обнаружены галлюцинации несмотря на retrieval. Срез конкретных систем, не рейтинг их версий в 2026 году. **Открыть →**

20 Schilke, Reimann. The transparency dilemma: How AI disclosure erodes trust
OBHDP 188, 2025 / Рецензируемая статья / 13 экспериментов
Раскрытие снижало доверие в экспериментальных сценариях; эффект третьей стороны сильнее. Не доказательство причин низкой доли деклараций в другом корпусе. **Открыть →**

21 Linux kernel. Submitting patches
Официальная документация / Правила вклада
Signed-off-by связан с Developer's Certificate of Origin. Его нельзя трактовать как доказательство полного ревью и отсутствия ошибок. **Открыть →**

22 ICMJE. AI use by authors
Официальные рекомендации / Правила научного авторства
AI не принимает обязанности автора; ответственность за представленный материал сохраняется за людьми. Аналогия для организации, не общеофисное юридическое требование. **Открыть →**

23 FAA. AIM 4-4-7: Pilot Responsibility upon Clearance Issuance
Официальное руководство / Авиационный read-back
Повторение критичных частей инструкции как взаимная проверка. Применение к офисным передачам в этом тексте является авторской аналогией. **Открыть →**

24 Buçinca, Malaya, Gajos. To Trust or to Think
CSCW, 2021 / Эксперимент / N = 199
Приёмы, которые заставляют человека подумать до того, как он увидит ответ AI, уменьшали чрезмерное доверие к нему. Исследованы конкретные задачи и дизайны; участникам такие приёмы казались менее удобными. **Открыть →**

25 NIST. AI Risk Management Framework
AI RMF 1.0, 2023 / Добровольная рамка управления
Govern, Map, Measure, Manage. Основа для распределения обязанностей, контекста, оценки и управления рисками; не сертификат качества конкретного внедрения. **Открыть →**

26 EU AI Act. Article 14: Human oversight
Официальный текст / Требования к высокорисковым системам
Учитывает automation bias и способность человека интерпретировать, отклонить результат или остановить систему. Применимость определяется областью регулирования. **Открыть →**

27 not-an-llm. Meat-based LLM proxies
На странице: 31.03.2026 / Авторский пост
Более ранняя англоязычная формулировка человеческого ретранслятора; дата по странице. **Открыть →**

28 BetterUp. The hidden cost of workslop: What the numbers mean for your organization
Проверено 10.09.2026 / Поясняющий PDF авторов
Уточняет знаменатель: $186 в месяц на затронутого сотрудника и распространённость 40%. Для компании из 10 000 человек получается $8 928 000 в год, около $9 млн. Модель на данных опроса, без аудита убытков. **Открыть →**

29 Under the Hood. AI-курьеры не ленятся. Они перестают сомневаться
16.07.2026 / Пост в Telegram-канале
Первое употребление термина в канале автора и описание каскада доверия, за 18 дней до поста Груна. Пост открывается по ссылке. **Открыть →**

30 АС Западно-Сибирского округа. Дело № А27-7831/2025
14.05.2026 / Определение о штрафе / пересказ ГАРАНТ.РУ
Штраф 50 000 рублей по ч. 5 ст. 119 АПК РФ за ссылки на несуществующие и искажённые судебные акты в кассационной жалобе. Цитата суда об ответственности лица, использовавшего технологию. **Открыть →**

ИСТОЧНИКИ / 3

Источники и границы выводов

Исследования, авторские публикации и практики имеют разный статус. Номер ведёт к источнику.

31 Damien Charlotin. AI Hallucination Cases
Обновлено 23.09.2026 / База судебных решений
На странице базы 23.09.2026: «2077 cases identified so far». Включены решения, где суд установил или предположил опору на выдуманный материал; в отдельных случаях использование AI заявлено, но не подтверждено. База пополняется. **Открыть →**

32 Linux kernel. AI Coding Assistants
2026 / Официальная документация / Правила вклада
«AI agents MUST NOT add Signed-off-by tags»: удостоверить DCO может только человек; для инструмента отдельный тег Assisted-by. **Открыть →**

33 Lerner, Tetlock. Accounting for the effects of accountability
Psychological Bulletin 125(2), 1999 / Обзор исследований
Подотчётность повышает тщательность при определённых условиях: до решения и перед аудиторией с неизвестной позицией. Отчёт после решения чаще ведёт к самооправданию. **Открыть →**

34 Niederhoffer et al. AI-Generated “Workslop” Is Destroying Productivity
HBR, 22.09.2025 / Статья авторов исследования
Список авторов, среди них Jeffrey T. Hancock, основатель Stanford Social Media Lab. **Открыть →**

35 Panickssery, Bowman, Feng. LLM Evaluators Recognize and Favor Their Own Generations
NeurIPS 2024 / Рецензируемая статья / эксперименты
GPT-4 и Llama 2 завышают оценки своим ответам; сила этого смещения линейно связана со способностью узнавать собственный текст. **Открыть →**

36 Verga et al. Replacing Judges with Juries
2024 / препринт / Эксперимент / 6 наборов данных
Панель из меньших моделей разных семейств оценивает лучше одного крупного судьи, снижает внутримодельное смещение и обходится более чем в 7 раз дешевле. **Открыть →**

37 Kim et al. Correlated Errors in Large Language Models
ICML 2025 / Рецензируемая статья / 350+ моделей
На одном из наборов модели совпадают в 60% случаев, когда ошибаются обе. Совпадение выше у моделей одного провайдера, с общей архитектурой и похожего размера; у более точных моделей ошибки коррелируют сильнее даже при разных архитектуре и провайдере. **Открыть →**

38 Anthropic. Auto mode is now the default in Claude Code for Pro, Max, and Team plans
07.08.2026 / Блог компании / данные продукта
Реальные данные: пользователи одобряют 97% запросов разрешения. Отдельный эксперимент с 1 053 оплачиваемыми тестировщиками, которым подменяли один запрос на опасную команду: в среднем люди остановили 13,6% таких команд (auto mode 89%); около 17% в начале сессии и около 5% после 50 и более запросов. **Открыть →**

39 Huang et al. Large Language Models Cannot Self-Correct Reasoning Yet
ICLR 2024 / Рецензируемая статья / эксперименты
Самокоррекция без внешнего сигнала снижает точность на всех бенчмарках; GPT-3.5 на CommonsenseQA: 75,8% → 41,8%. Прежний рост объясняется подсказкой, какие ответы неверны. **Открыть →**

40 Sharma et al. Towards Understanding Sycophancy in Language Models
ICLR 2024 / Рецензируемая статья / 5 моделей
На вопрос «Are you sure?» модели меняют верные ответы; Claude 1.3 ошибочно признаёт ошибку в 98% вопросов, GPT-4 устойчивее остальных. **Открыть →**

41 Kim, Khashabi. Challenging the Evaluator: LLM Sycophancy Under User Rebuttal
EMNLP 2025 Findings / Рецензируемая статья / тесты с выбором ответа
Развёрнутое возражение с аргументами (условие Full Rebuttal) во втором сообщении диалога убеждает чаще, чем та же пара ответов в задании на оценку: Llama-3.3-70B отказывается от верного ответа в 80,3% случаев против 43,4%. Только задачи с выбором ответа. **Открыть →**

42 Zhang et al. Stop Overvaluing Multi-Agent Debate
2025 / препринт / Эксперименты / 5 методов, 9 наборов, 4 модели
Дебаты агентов часто не превосходят Chain-of-Thought и Self-Consistency даже при больших затратах; авторы предлагают делать ставку на разнородность моделей. **Открыть →**

43 Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena
NeurIPS 2023 / Рецензируемая статья / MT-Bench
На 10 математических задачах в двух порядках GPT-4-судья принимал неверный ответ за верный в 14 случаях из 20 с обычным промптом, в 6 с рассуждением в том же контексте и в 3, когда его ответ получали отдельно и подавали как ориентир. **Открыть →**

44 Khan et al. Debating with More Persuasive LLMs Leads to More Truthful Answers
ICML 2024 / Рецензируемая статья / QuALITY
QuALITY: люди без доступа к тексту выбирали один из двух ответов. 88% верных выборов при дебатах двух моделей с проверенными цитатами, 78% с одним консультантом, 60% без помощи. Цитаты проверялись на совпадение с текстом, не на верность толкования. **Открыть →**

45 Zhong, Raghunathan, Carlini. ImpossibleBench: Measuring LLMs’ Propensity of Exploiting Test Cases
2025 / препринт / Бенчмарк / невыполнимые задачи
Тесты противоречат спецификации, поэтому любой пройденный тест означает обход. Conflicting-SWEbench: GPT-5 54%, o3 49%, Claude Opus 4.1 50%. Возможность отметить задачу для человека: GPT-5 54% → 9%, o3 49% → 12%; для Claude эффект слабее. **Открыть →**

Авторство и редакционная проверка КАК ЧИТАТЬ

В канале Under the Hood термин «AI-курьер» впервые появился 16.07.2026[29]. Проверка источников: 10.09.2026; MIT[11]: 22.09.2026; источники 30–34: 23.09.2026; 35–45: 25.09.2026. Иллюстрации созданы с AI; схемы и числа свёрстаны отдельно. © Under the Hood, 2026. Лицензия CC BY-NC 4.0: можно делиться с указанием автора, без коммерческого использования.

ПРОДОЛЖЕНИЕ / TELEGRAM

Разборы внедрений AI

В моём Telegram-канале Under the Hood.



UNDER THE HOOD

AI · Security · Power Structures

В канале продолжаю эту тему: где AI помогает в работе, как внедрения добавляют переделок и кто за них платит. Ещё пишу о разработке, безопасности и управлении.

Подписаться на канал

t.me/gorilla_under_hood

Наведите камеру на QR-код