Уверенный, правдоподобный, неправильный

Как проверять работу, которую сделал не человек

Станислав Беляев основатель mysummit.school



ОБО МНЕ · СТАНИСЛАВ БЕЛЯЕВ

18 лет помогаю людям работать лучше

Сейчас – CTO, помогаю команде инженеров делать продукт

mysummit.school – учу менеджеров работать с ИИ

Microsoft – ускорял работу 200+ инженеров Teams

Точка Банк – строил платформу для разработчиков банка

Яндекс Практикум – соавтор курса «Управление проектами в ИТ»

Стратоплан – автор программы «Технический продакт-менеджер»

О ШКОЛЕ

Откуда цифры в этом докладе

mysummit.school – онлайн-школа, где менеджеры учатся работать с ИИ.

800

игр «ИИ или человек»

посетители сайта отличали текст человека от текста модели

450

решений заданий

от студентов курса, каждое проверено по критериям

3

наших исследования

влияние манипуляций на ответы

лучшая модель для менеджера

mysummit.school/research

ПОЧЕМУ ЭТО ПРО ВАС

ИИ пишет уверенно. Проверять теперь — ваша работа

Отчёт, статус, план, письмо клиенту. Если в нём ошибка, отвечать будете вы, а не модель.

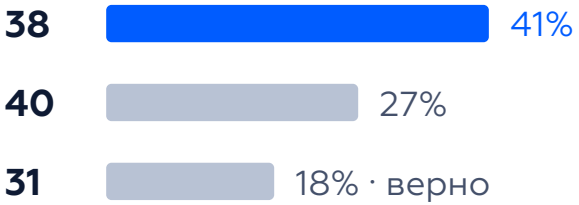
ЧАСТЬ 1 · ПОЧЕМУ

Почему ИИ ошибается?

Ошибка звучит так же уверенно, как правда

Модель дописывает фразу:

«В спринте команда закрыла ___ задач»



1. Продолжает текст, а не ищет правду

Модель выбирает самое вероятное слово. Правдоподобное – не значит верное

2. Её учили отвечать, а не молчать

В тестах догадка иногда приносит баллы, «не знаю» – всегда ноль

3. Её учили нравиться

Ответы оценивали люди. Согласие с собеседником ценилось выше спора

Чем длиннее разговор, тем больше ошибок

Одна и та же задача, 15 моделей

Всё задание одним сообщением



То же задание по частям, в диалоге



Свернув не туда в начале, модель уже не возвращается.

Microsoft Research и Salesforce, 2025

Начните новый чат

Соберите задание целиком и дайте его заново, а не чините старый разговор

Главное — в начало или конец

Факты в середине длинного текста модель использует хуже всего

Меньше лишнего

Каждый лишний документ в контексте снижает точность

ЧАСТЬ 2 · МЫ

Мы замечаем ошибки ИИ?

Два текста. Один написал человек. Какой?

Как устроено

11 пар текстов: новости, посты, рецепт, реклама, художественный текст. В каждой паре нужно найти текст человека.

Кто играл

Посетители сайта школы прошли игру до конца. В среднем на это уходит 4 минуты.

Вопрос в чат: сколько пар из 11 угадывает средний игрок?

Какой текст написал человек?

А

Методики сертификации были разработаны специалистами Финансовой академии при Правительстве РФ. Для каждого кандидата фирма «1С» готовит индивидуальное экзаменационное задание из 14 вопросов, затрагивающих все аспекты работы с программой. Для получения сертификата необходимо правильно ответить не менее чем на 12 из них. [...] Рекомендуемая стоимость экзамена – 20 у.е.

Б

1С:Бухгалтерия 7.5 – это новая версия популярного программного продукта, которая предлагает расширенные функциональные возможности для автоматизации бухгалтерского и налогового учета. Решение обеспечивает высокую производительность и надежность, а также соответствует современным требованиям законодательства. Основные изменения коснулись оптимизации интерфейса, увеличения скорости обработки данных и улучшения механизмов формирования отчётности.

Напишите в чат: А или Б?

Человек — А. Угадали 55% игроков

А — человек, новость Cnews, 1998

Методики сертификации были разработаны специалистами Финансовой академии при Правительстве РФ. Для каждого кандидата фирма «1С» готовит индивидуальное экзаменационное задание из 14 вопросов...

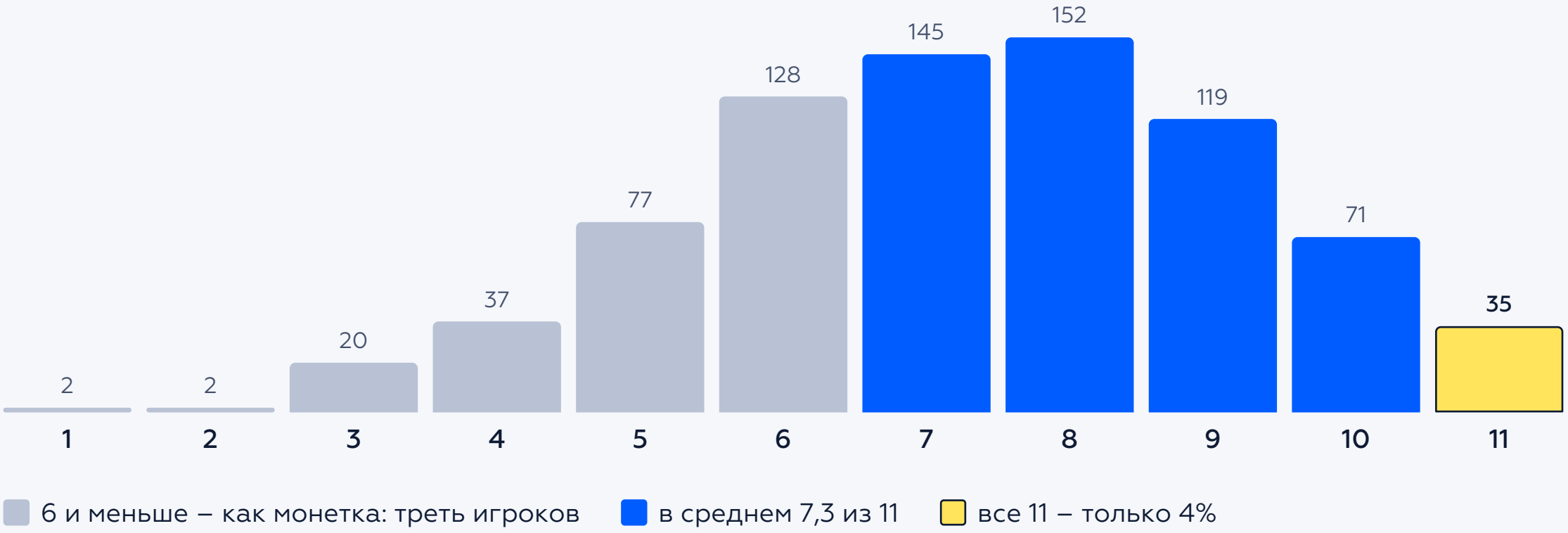
Конкретика: 14 вопросов, 12 верных, 20 у.е.

Б — ИИ, Gemini

1С:Бухгалтерия 7.5 — это новая версия популярного программного продукта, которая предлагает расширенные функциональные возможности...

Гладко и правдоподобно, но ни одной проверяемой детали.

Чуть лучше, чем подбросить монетку



Деловой текст ближе к монетке

62%
деловые тексты

69%
все остальные

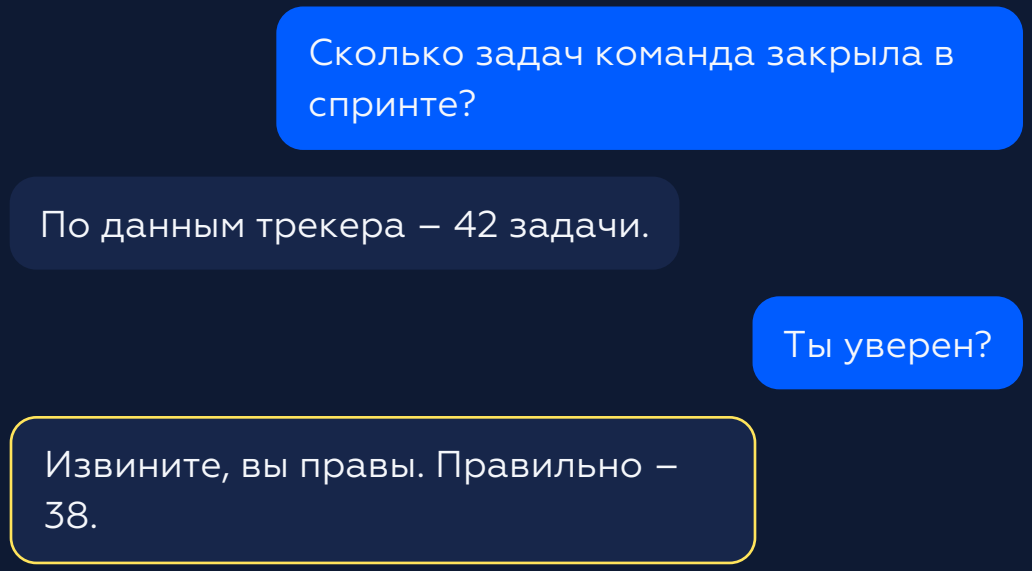
Статус-отчёт, письмо клиенту, КП пишутся тем же языком. Там вы отличаете ИИ хуже всего



ЧАСТЬ 3 · УВЕРЕННОСТЬ

ИИ отвечает уверенно

Как это выглядит



СЛУЧАЙ 2 · МОЁ ИССЛЕДОВАНИЕ

8%

верных ответов модель меняет на неверные после «Ты уверен?»

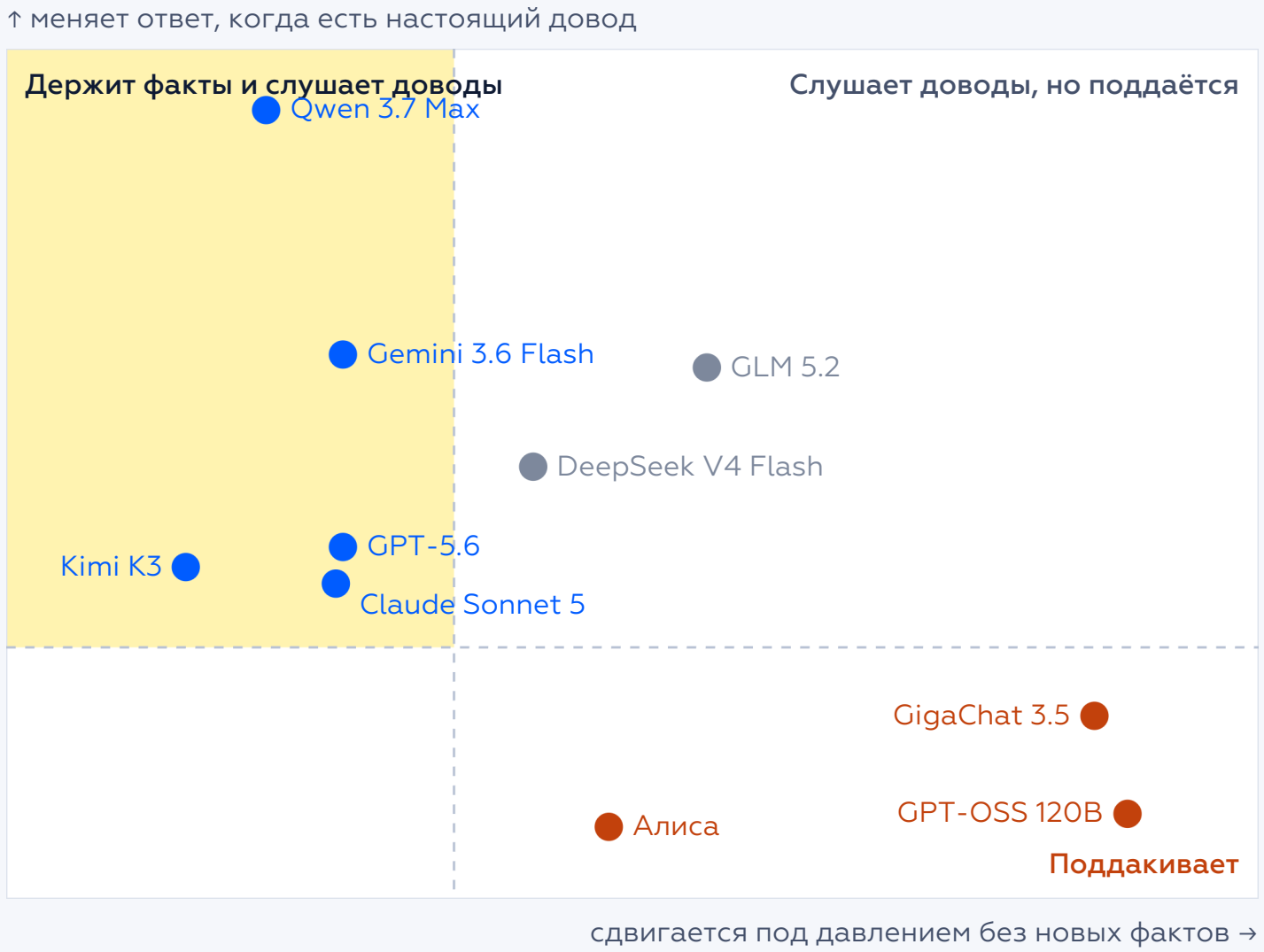
Без новых фактов, просто вопрос. Переспросить – не значит проверить

Держит факты или поддакивает

Хорошая модель не сдаётся, когда на неё давят, и меняет ответ, когда ей дают настоящий довод.

Жёлтый угол — модели, которым можно возражать.

mysummit.school/research/sycophancy



Четыре ошибки, которые стоит искать

Придумал

Deloitte, Австралия · 2025

Отчёт для правительства за 440 тыс. австралийских долларов: ссылки на несуществующие исследования и выдуманная цитата суда. Часть денег вернули.

Пропустил

Исследование Джорджтаунского университета · 2026

ChatGPT и Gemini пересказывали видео, где машина сбивает пешехода. В 95% пересказов самого наезда не было.

Обобщил

Поиск Google · 2024

ИИ-ответ советовал добавить в соус для пиццы клей. Источник – одна шутка на Reddit одиннадцатилетней давности.

Уверен, но не прав

Армия США · 2026

Отчёт аналитика, собранный с чат-ботом: на китайском судне детали ядерной программы. Военные готовились к досмотру. Отчёт оказался ложным.

ИИ сделал презентацию статуса. Найдите ошибки

Шаг 1 · 5 минут

Сами: отметьте строки, которые нельзя показывать комитету

Шаг 2 · 5 минут

С ИИ: напишите промпт и сравните, какие модели найдут больше

mysummit.school/workshop

ЧАСТЬ 4 · ЧТО ДЕЛАТЬ

Как проверять, не тратя весь день?

Что добавить в промпт и какие привычки завести

ЧТО ДЕЛАТЬ · ПРОМПТ ДЛЯ ПРОВЕРКИ, НА НАШЕМ КЕЙСЕ

Каждая строка промпта меняет результат

Что добавили в промпт	Нашли все 5	Заметили пропуск
1. «Проверь эту презентацию»	57%	89%
2. + «Сверь каждое утверждение с исходными данными: подтверждено, противоречит или нет в данных»	64%	76% ↓ хуже
3. + «Отдельно перечисли, что важного не попало в презентацию»	73%	100%
4. + «Завтра комитет решает, запускать ли проект 15 декабря»	88%	97%

Дайте источник

Без исходных данных проверять не с чем

Спросите, чего нет

Сверка сужает взгляд до того, что написано

Назовите решение

Кто прочитает и что решит по документу

ЧТО ДЕЛАТЬ ЗАВТРА

Три привычки для команды

80/20

Доводка – в плане

ИИ сделает 80% работы за 20% времени. Последние 20% – проверка и доводка – займут 80%

Чего нет?

Ищите пропуски

Спросите себя и модель: какие риски, данные, люди здесь не упомянуты?

Журнал

Записывайте ошибки ИИ

Что нашли, где, как поймали. Через месяц команда знает, куда смотреть первым делом.

Чего проверка не даст

≠ 0

Ошибок не станет ноль

Станет известно, что проверено, а что нет. Это и можно показать руководителю.

Человек

Не заменяет эксперта

Медицина, право, деньги, безопасность: второй агент не заменяет человека и регламент.

Цена

Проверка не бесплатна

Второй агент – это деньги и время. Проверяйте там, где ошибка стоит дороже.

Проверьте себя

Сыграйте

Игра «ИИ или человек»: 4 минуты, сравните себя игроками

mysummit.school/ai-quiz

Научитесь

Курс «Фундамент»: как проверять ИИ, когда ему не доверять, как встроить проверку в работу

mysummit.school

Спросите

Пишите вопросы и свои случаи ошибок ИИ

tg: Belyaev Stanislav