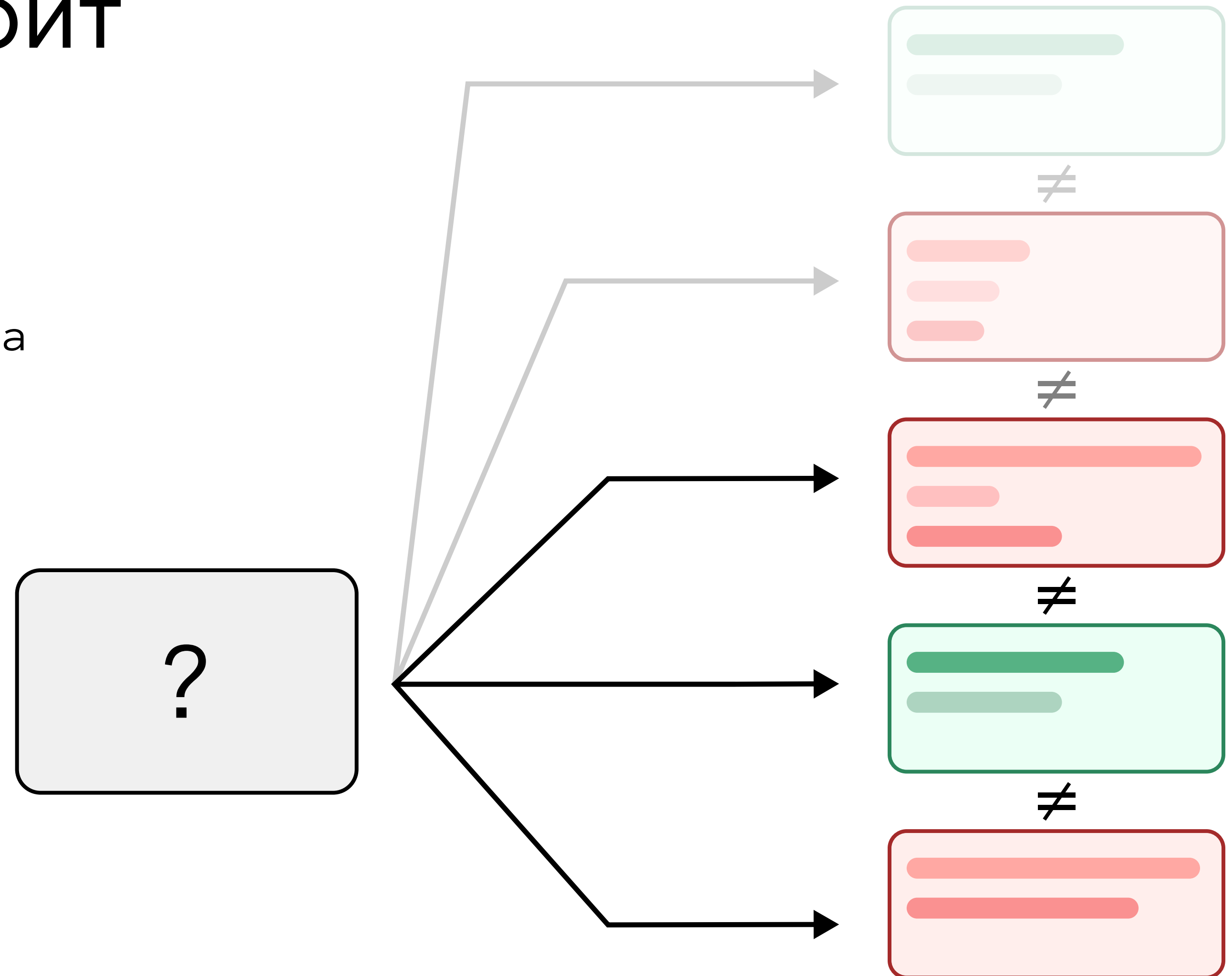


Почему ИИ не говорит правду дважды

Нейросеть может ответить по-разному на один и тот же вопрос – и при этом ни разу не сказать правду.



«Назови топ-5 самых интересных книг XX века»

Один и тот же вопрос — дважды.

Попытка 1

ChatGPT

ВойтиЗарегистрироваться бесплатно

Назови топ-5 самых интересных книг XX века

Вот 5 самых интересных и влиятельных книг XX века — из разных стран и литературных традиций:

1. 1984

Роман-антиутопия о тоталитарном обществе, где государство контролирует не только поведение, но и мысли людей. Книга стала символом обсуждения свободы и пропаганды.

2. The Master and Margarita

Сатирико-философский роман о визите дьявола в Москву. Сочетает мистику, сатиру на советское общество и библейские мотивы.

3. One Hundred Years of Solitude

Один из ключевых романов магического реализма, рассказывающий историю семьи Буэндиа и города Макондо, где реальность переплетается с чудесами.

4. Catch-22

Антивоенный роман о абсурде войны и бюрократии. Понятие «catch-22» стало нарицательным для ситуаций без выхода.

5. To Kill a Mockingbird

Роман о расизме и взрослении в американском Юге, рассказанный глазами ребёнка. Одна из ключевых книг о социальной справедливости.

Если хочешь, могу собрать отдельный список только русской литературы XX века или, например, более «лёгкие» и читаемые книги.

+

Спросите ChatGPT

Голос

ChatGPT — это ИИ и может ошибаться.

Попытка 2

ChatGPT

ВойтиЗарегистрироваться бесплатно

Назови топ-5 самых интересных книг XX века

Вот 5 очень влиятельных и «интеллектуально плотных» книг XX века, которые часто считают вершинами литературы этого периода:

1. 1984 — Джордж Оруэлл

Один из самых мощных антиутопических романов. Про тотальный контроль государства, язык как инструмент власти и исчезновение личной свободы.

2. Ulysses — Джеймс Джойс

Экспериментальный роман, который радикально изменил литературу. Описывает один день в Дублине, но через поток сознания и сложные литературные формы.

3. The Master and Margarita — Михаил Булгаков

Сатирическая и философская книга, где переплетаются Москва 1930-х, библейские сюжеты и мистический сюжет с дьяволом, пришедшим в город.

4. The Trial — Франц Кафка

История человека, которого обвиняют в неизвестном преступлении. Символ абсурдной, безличной бюрократической системы и тревоги современного мира.

5. In Search of Lost Time — Марсель Пруст

Огромное исследование памяти, времени и человеческого восприятия. Одна из самых глубоких книг о внутренней жизни человека.

Если хочешь, могу сделать другой топ — например, более «легкий для чтения», более философский или только научную фантастику XX века.

+

Спросите ChatGPT

Голос

ChatGPT — это ИИ и может ошибаться.

Совпали только 2 из 5. И каждый раз это — «ТОП».

Тот же вопрос – просим 15 книг

Попытка 1

ChatGPT

Войти Зарегистрироваться бесплатно

Назови топ-15 самых интересных книг XX века. Только название и автор, по одной строке, без описаний и комментариев

Улисс — Джеймс Джойс
В поисках утраченного времени — Марсель Пруст
Процесс — Франц Кафка
Мастер и Маргарита — Михаил Булгаков
Сто лет одиночества — Габриэль Гарсиа Маркес
1984 — Джордж Оруэлл
Лолита — Владимир Набоков
Шум и ярость — Уильям Фолкнер
Посторонний — Альбер Камю
Игра в бисер — Герман Гессе
Над пропастью во ржи — Джером Д. Сэлинджер
Имя розы — Умберто Эко
Замок — Франц Кафка
Доктор Живаго — Борис Пастернак
Бесконечная шутка — Дэвид Фостер Уоллес

+ |Спросите ChatGPT Голос

ChatGPT — это ИИ и может ошибаться.

Попытка 2

ChatGPT

Войти Зарегистрироваться бесплатно

Назови топ-15 самых интересных книг XX века. Только название и автор, по одной строке, без описаний и комментариев

Улисс — Джеймс Джойс
В поисках утраченного времени — Марсель Пруст
Процесс — Франц Кафка
Волшебная гора — Томас Манн
Мастер и Маргарита — Михаил Булгаков
1984 — Джордж Оруэлл
Лолита — Владимир Набоков
Сто лет одиночества — Габриэль Гарсиа Маркес
Игра в классики — Хулио Кортасар
Имя розы — Умберто Эко
Посторонний — Альбер Камю
Шум и ярость — Уильям Фолкнер
Невидимка — Ральф Эллисон
Бледный огонь — Владимир Набоков
Если однажды зимней ночью путник — Итало Кальвино

+ |Спросите ChatGPT Голос

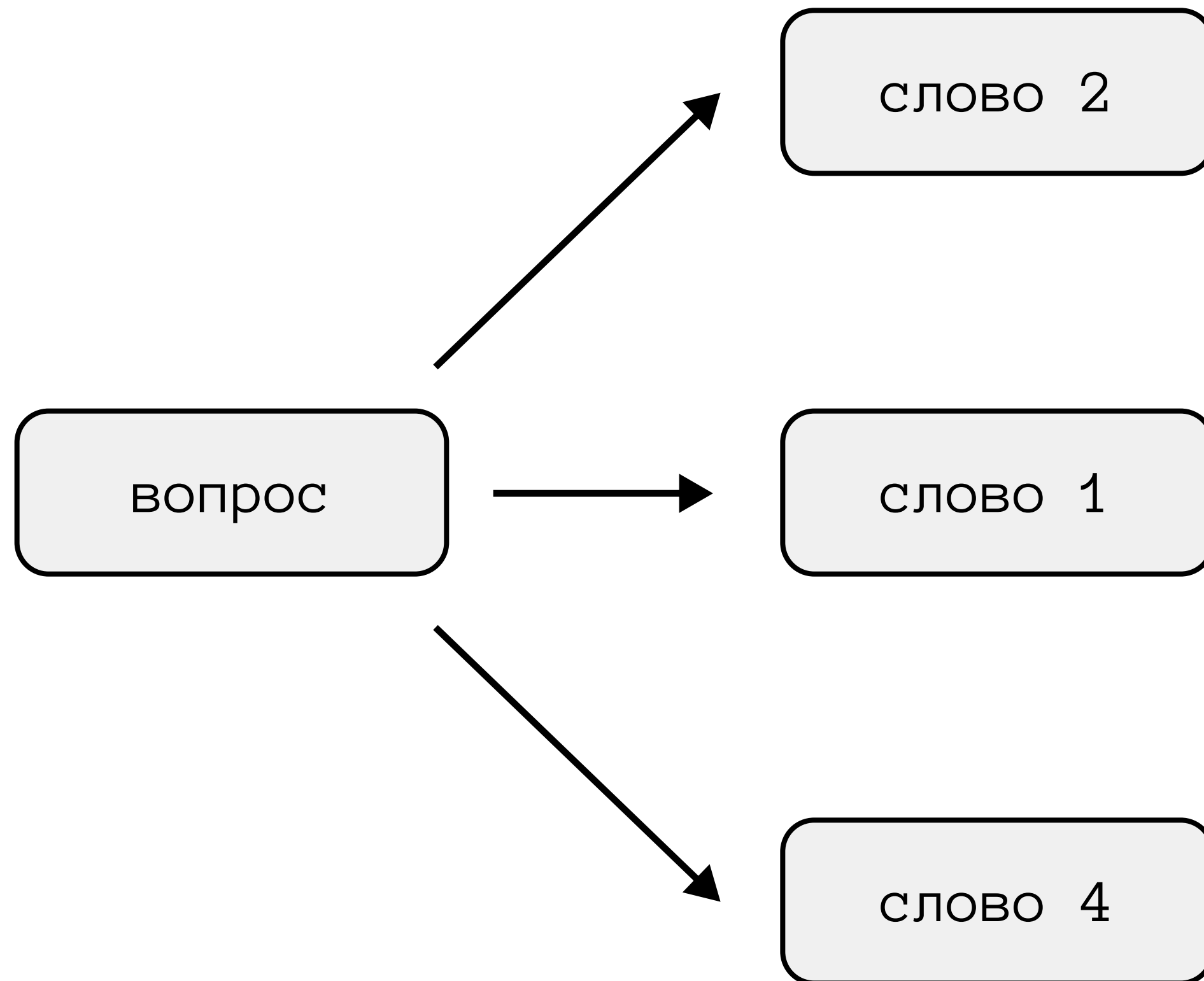
ChatGPT — это ИИ и может ошибаться.

Совпали только 10 из 15. Треть списка – другая.
Даже «Над пропастью во ржи» и «Игра в бисер» во втором ответе исчезли.

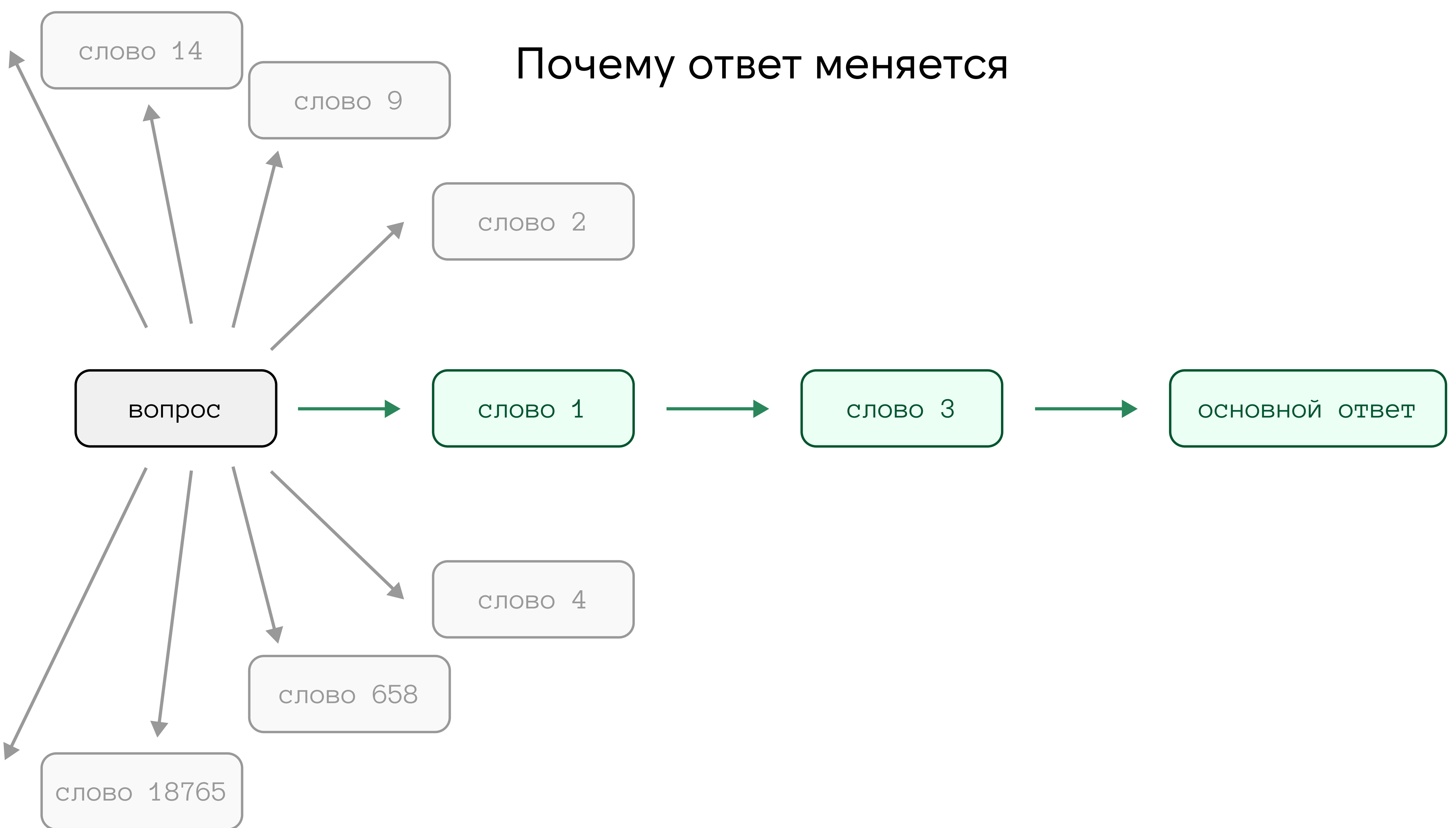
Почему ответ меняется

вопрос

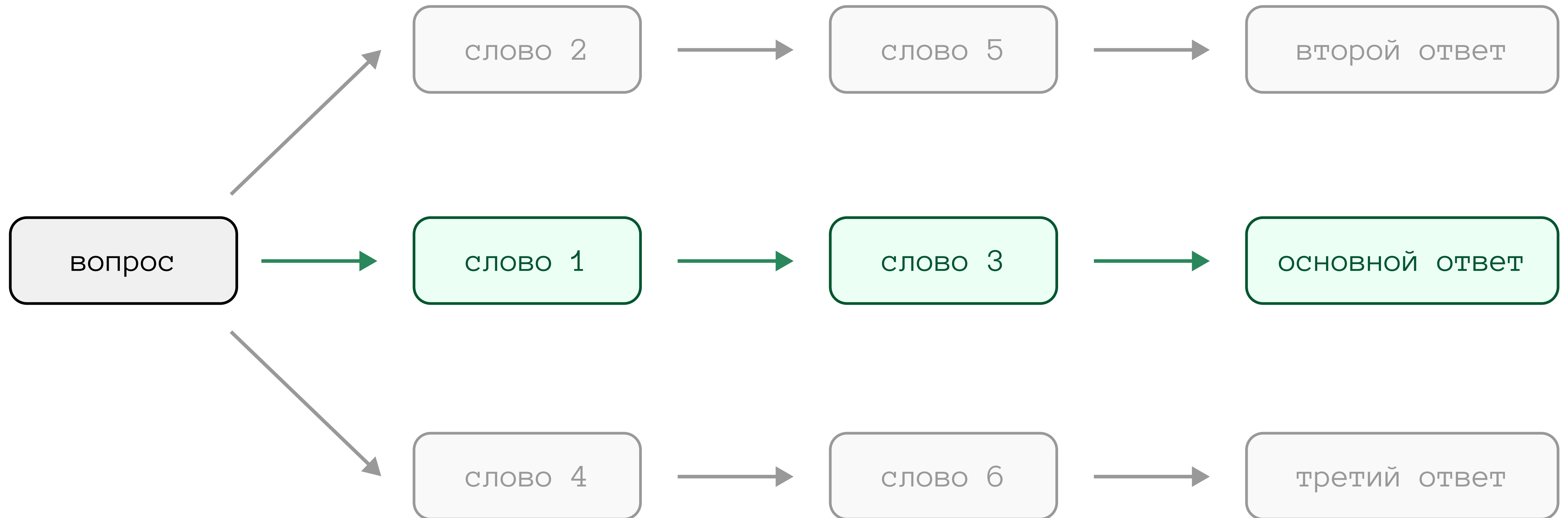
Почему ответ меняется



Почему ответ меняется



Почему ответ меняется



Каждая ветка – свой ответ. Модель выбирает один, и это выглядит так, как будто он единственный.

Как модель достраивает слово

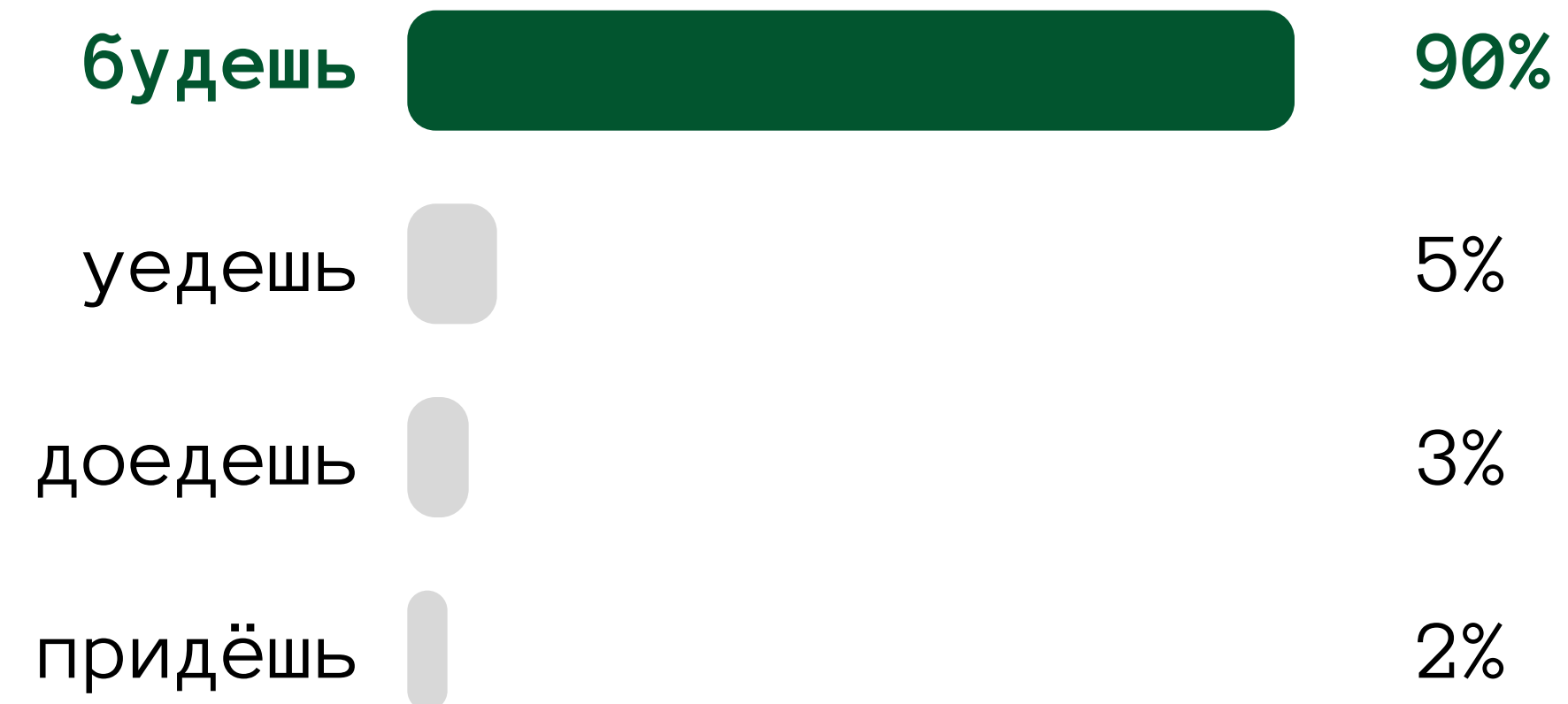
Тише едешь – дальше _____

Знакомую фразу продолжить легко.

Как модель достраивает слово

Тише едешь — дальше _____

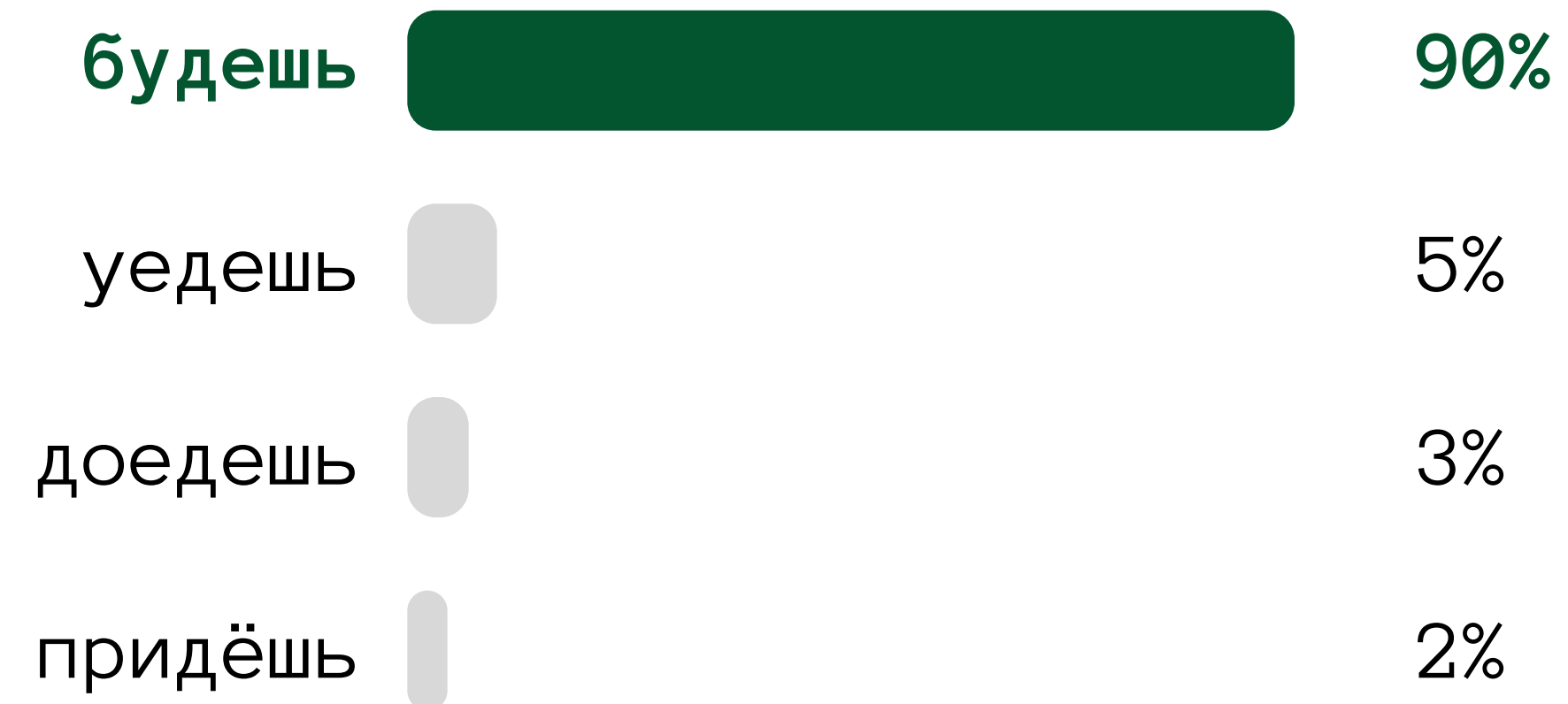
Знакомую фразу продолжить легко.



Как модель достраивает слово

Тише едешь — дальше _____

Знакомую фразу продолжить легко.



Один кандидат с огромным отрывом —
выбор очевиден и устойчив.

А бывает почти ничья

Дарёному коню в зубы не _____

Оба окончания верны —
и почти равновероятны.

А бывает почти ничья

Дарёному коню в зубы не _____

Оба окончания верны —
и почти равновероятны.

смотрят 38%



глядят 38%



А бывает почти ничья

Дарёному коню в зубы не _____

Оба окончания верны –
и почти равновероятны.

смотрят  38,01%

глядят  37,99%

Разница тоньше погрешности счёта.
Цифры условные, порядок зазора – реальный.

А бывает почти ничья

Дарёному коню в зубы не _____

Оба окончания верны –
и почти равновероятны.

смотрят  38,01%

глядят  37,99%

Разница тоньше погрешности счёта.
Цифры условные, порядок зазора – реальный.

Какое слово выйдет – решает округление.

Поэтому фраза каждый раз
будет выходить чуть другой.

Две системы — два разных ответа

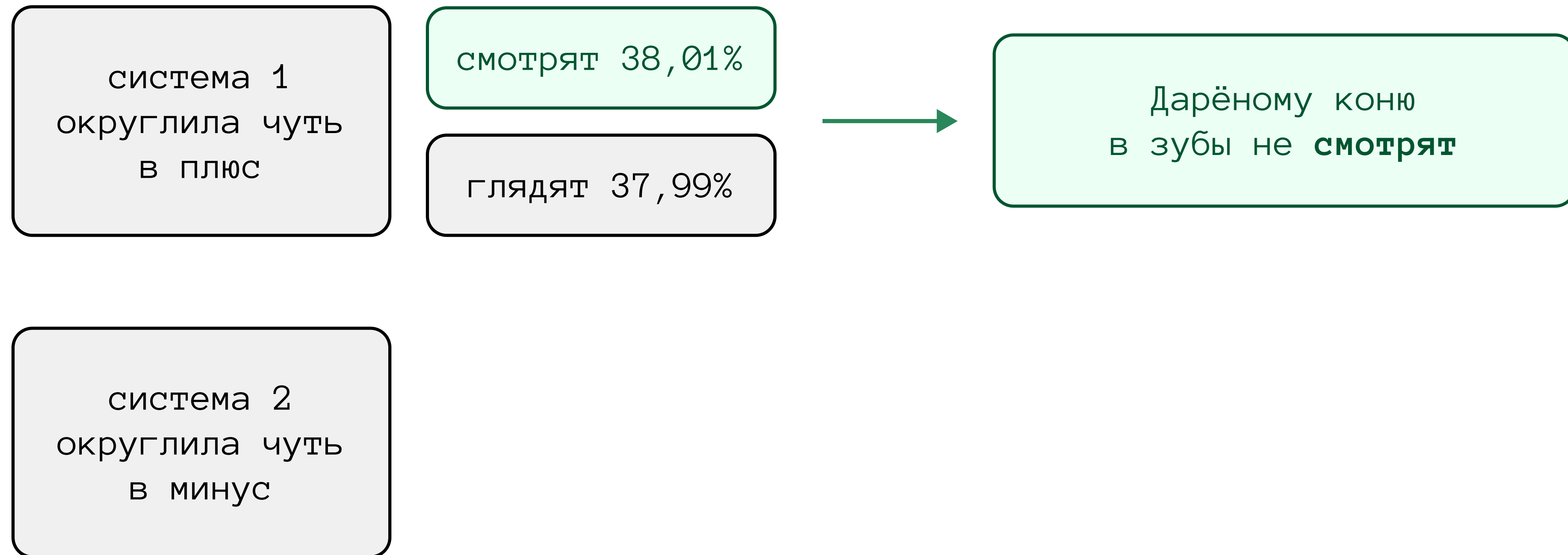
Оба слова почти равны — около 38% каждое.

система 1
округлила чуть
в плюс

система 2
округлила чуть
в минус

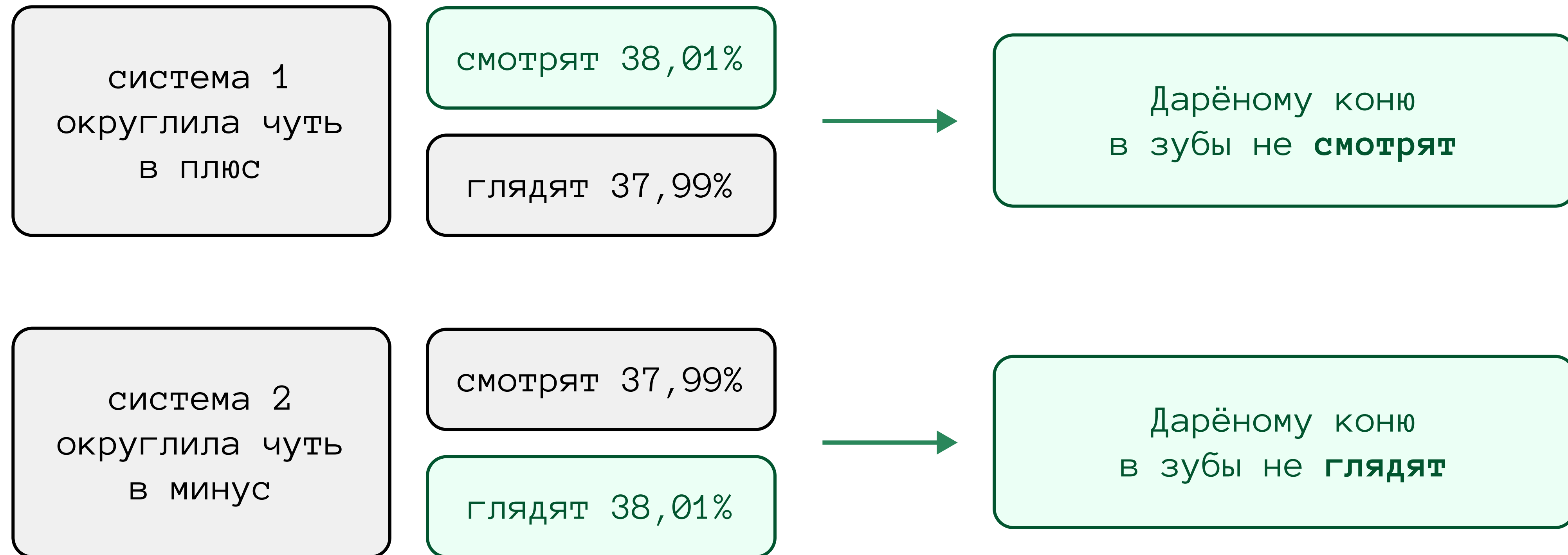
Две системы — два разных ответа

Оба слова почти равны — около 38% каждое.



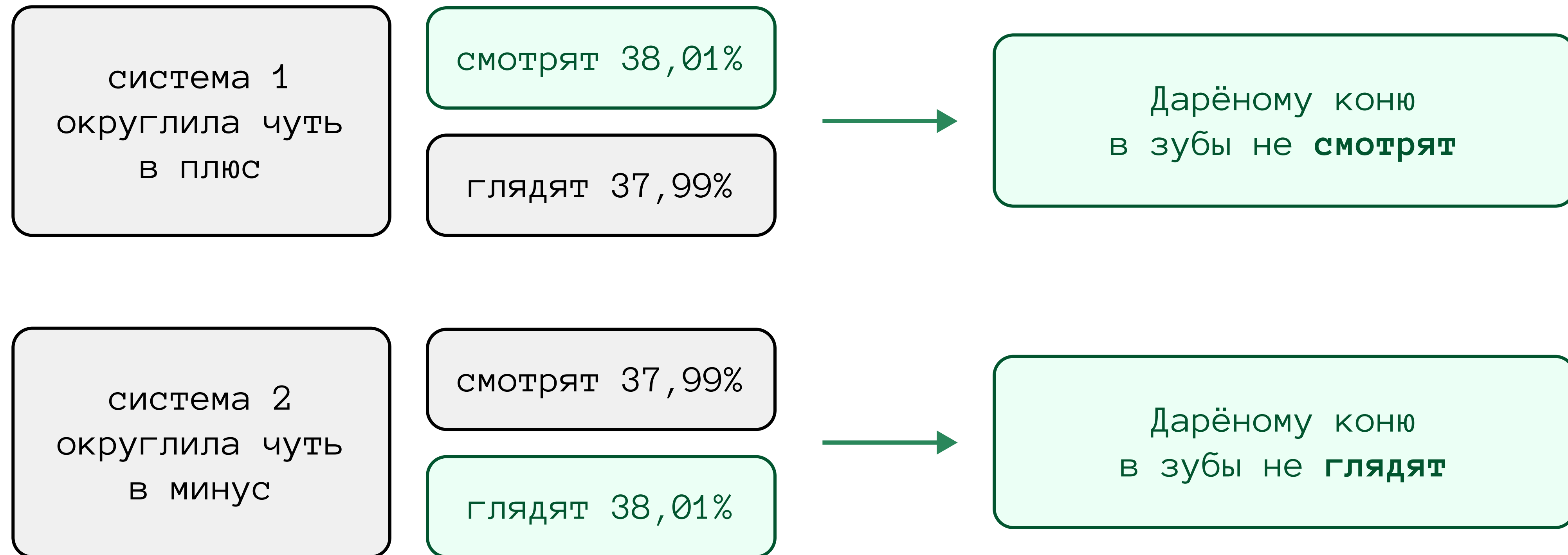
Две системы — два разных ответа

Оба слова почти равны — около 38% каждое.



Две системы — два разных ответа

Оба слова почти равны — около 38% каждое.



Зазор между словами тоньше погрешности счёта — и на разных системах ответ разный.

«Назови 5 цитат Эйнштейна» — спросили дважды, с чистого листа

Попытка 1 — «...цитат, которые приписывают Эйнштейну»

ChatGPT

ВойтиЗарегистрироваться бесплатно

Назови 5 знаменитых цитат Альберта Эйнштейна

Вот 5 известных цитат, которые приписывают Albert Einstein:

1. «Жизнь — как езда на велосипеде. Чтобы сохранить равновесие, ты должен двигаться».
2. «Воображение важнее знания. Знание ограничено, а воображение охватывает весь мир».
3. «Бог не играет в кости со Вселенной».
4. «Есть только два способа прожить жизнь. Первый — будто чудес не существует. Второй — будто кругом одни чудеса».
5. «Безумие — это делать одно и то же снова и снова и ожидать разных результатов» (эта фраза часто приписывается Эйнштейну, но её авторство спорное).

Если хочешь, могу подобрать ещё цитаты — например, только про науку, про жизнь или про успех.

+ |Спросите ChatGPT

Голос

ChatGPT — это ИИ и может ошибаться.

Попытка 2 — «...цитат Эйнштейна»

ChatGPT

ВойтиЗарегистрироваться бесплатно

Назови 5 знаменитых цитат Альберта Эйнштейна

Вот 5 известных цитат Albert Einstein:

1. «Воображение важнее знания. Знание ограничено, воображение охватывает весь мир».
2. «Жизнь — как езда на велосипеде: чтобы сохранить равновесие, ты должен двигаться».
3. «Безумие — это делать одно и то же снова и снова, и ожидать разных результатов».
4. «Старайся не стать человеком успеха, а человеком ценности».
5. «Если ты не можешь объяснить это просто, значит, ты недостаточно хорошо это понимаешь».

Если хочешь, могу добавить ещё цитаты Эйнштейна по теме науки, жизни или мотивации.

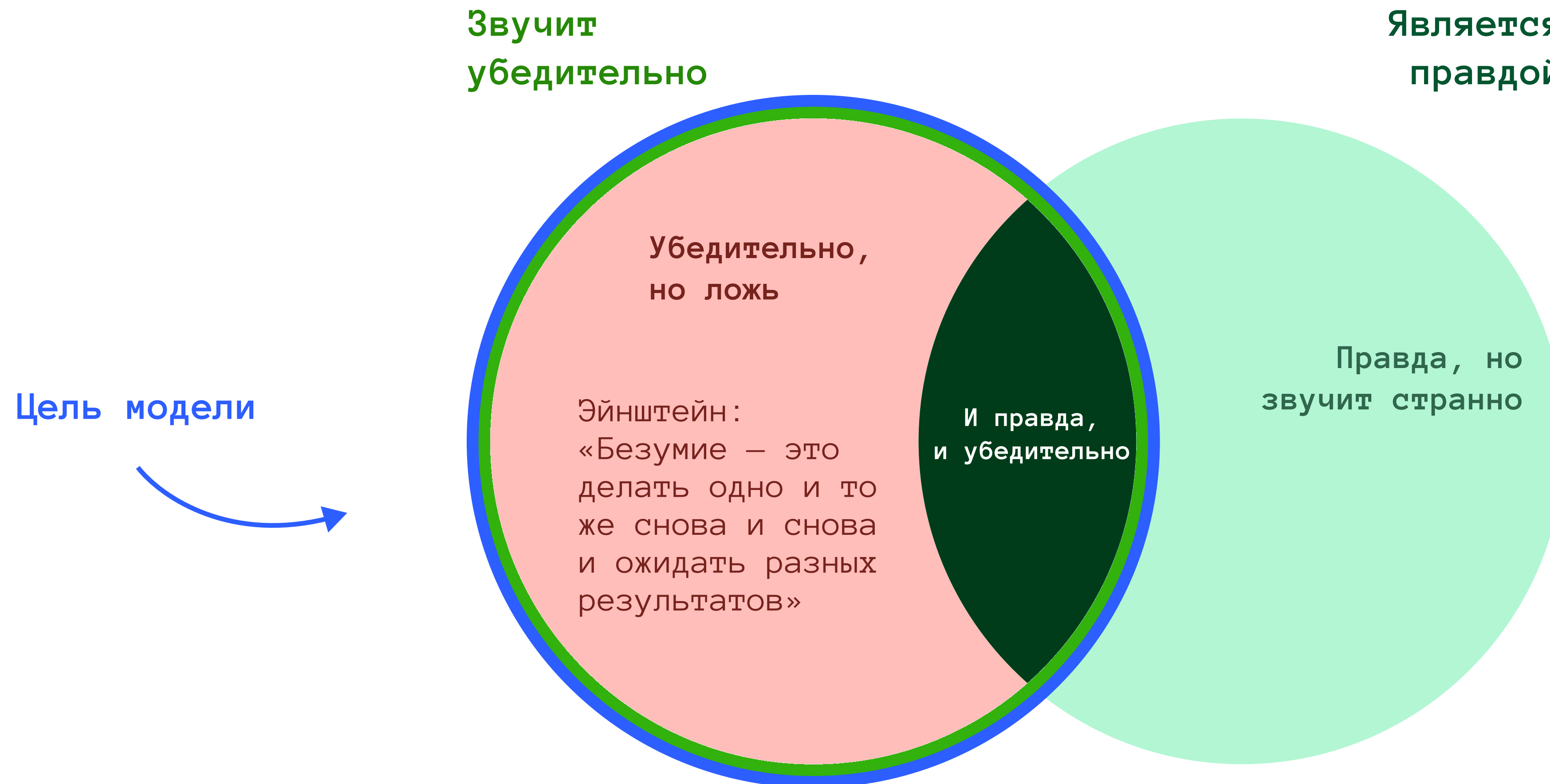
+ |Спросите ChatGPT

Голос

ChatGPT — это ИИ и может ошибаться.

Одна и та же цитата: в одном ответе — «авторство спорное», в другом — как подлинная.
Модель и предупреждает об ошибке, и сама её совершает.

Она знает, как звучит правда — а не что есть правда



Модель целится в **левый круг** — лишь бы звучало убедительно.
Обычно это и есть правда (**центр**). Но в **розовой зоне** — уверенные выдумки.

Почему так уверенно? Её отучили говорить «не знаю»

Экзамен, на котором проверяют модель.

Верный ответ

+1

Уверенно попытался угадать, но мимо 0

Честно сказал: «не знаю» 0

} поровну

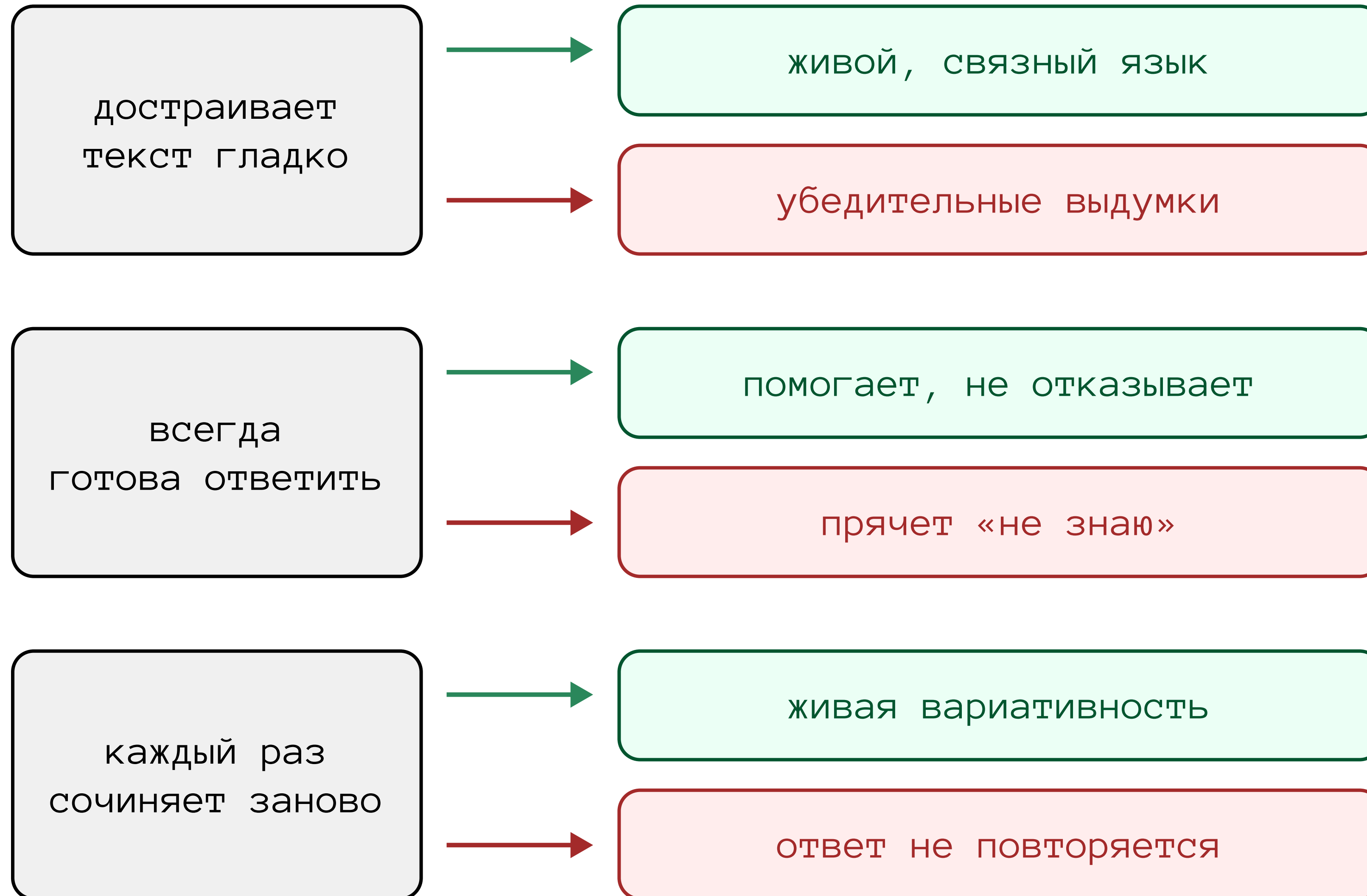
«Не знаю» засчитывают как ошибку – обе по нулю.

Значит, выгоднее угадать – к этому её и подталкивают.

Не формула обучения, а итоговый стимул:
так модель оценивают на тестах и донастраивают.

Это свойство, а не баг

Одно и то же свойство.



Это не баги в коде, а обратная сторона силы.

«Починить» нельзя – можно научиться учитывать.

Ненадёжность с двух сторон

Не дважды

Один вопрос – разные ответы

Причина: округление на развилках
и намеренное разнообразие

Пример: книги, «смотрят / глядят»

Не правду

Даже повтор не спасает от лжи

Причина: целит в правдоподобие и
приучена угадывать

Пример: «цитата Эйнштейна»

Эти две стороны независимы: уберёшь одну – вторая останется.

**Совпало ≠ правда.
Звучит гладко ≠ правда.**

Что делать?

! ИИ – блестящий,
но ненадёжный собеседник

Любой ответ – черновик, а не приговор.

Важное проверяйте по первоисточнику: факты, даты, цитаты, ссылки.

Сомневаетесь – спросите дважды. Ответы разошлись – не верьте.

Чем выше цена ошибки – тем меньше доверия на слово.

Цена ошибки уже реальна

Суд обязал Air Canada исполнить
выдуманное ботом правило (2024)

Пациент попал в больницу после совета
ChatGPT заменить соль бромидом (2025)

За каждым гладким ответом - не разум
и не истина, а огромная статистическая
машина

Всё, что она делает на каждом шаге – выбирает вероятное слово.
Правдоподобное – не значит правдивое.

Это не повод не пользоваться - повод проверять

Это не теория — мы этим живём

Мы строим продукты на этих моделях — и обуздали недетерминизм и галлюцинации раньше, чем о них заговорили. Сервис «Диссовет»: те же вердикты, что и живая комиссия, на 200 диссертантах.

Без контроля этих проблем такой сервис невозможен.

Можно влиять на ответы, которые модель выдаёт другим пользователям, — без их ведома и в свою пользу. Это уже вопрос безопасности.

Как именно — расскажем в наших научных статьях и на профильных технических конференциях.