Федеральный центр экспертизы в сфере безопасности
и развития искусственного интеллекта имени Д. А. Поспелова

Повторяемость ответов языковых моделей и стоимость правильного результата

Как отличить повторяемый ответ от правильного и сопоставить стоимость моделей с результатом, пригодным для дальнейшей работы.

Испытания моделей — 11 ноября 2026 года

Выбор языковой модели для обработки документов требует понять, за что платит организация: за каждый полученный ответ или за результат, который можно использовать после проверки. Повторный запрос тоже расходует средства, а совпадение двух ответов еще оставляет вопрос об их правильности. Центр сопоставил повторы, перефразирование и фактически зарегистрированные расходы, чтобы разделить эти основания выбора.

Практический вывод — повторяемость следует проверять вместе с правильностью, а стоимость считать для принятого результата. Qwen воспроизвел статус и значение во всех 96 парах повторов; в 21 паре оба ответа одинаково расходились с эталоном. Поэтому само совпадение ответов дает ограниченное основание для автоматического приема.

Работа опирается на общий корпус из 96 синтетических заданий для шести моделей серии CP-RU-DOC-2026-01. Учтены расходы интерфейсов и наблюдаемое поведение конкретных сервисов. Для расчета стоимости рабочего процесса организации потребуются также затраты на проверку источников, исправление ошибок и работу специалистов.

1. Предмет исследования

2. Корпус и устройство сравнения

3. Два уровня оценки формата

4. Совпадение, правильность и ссылки

5. Уверенность и передача на другую модель

6. Учет расходов и времени

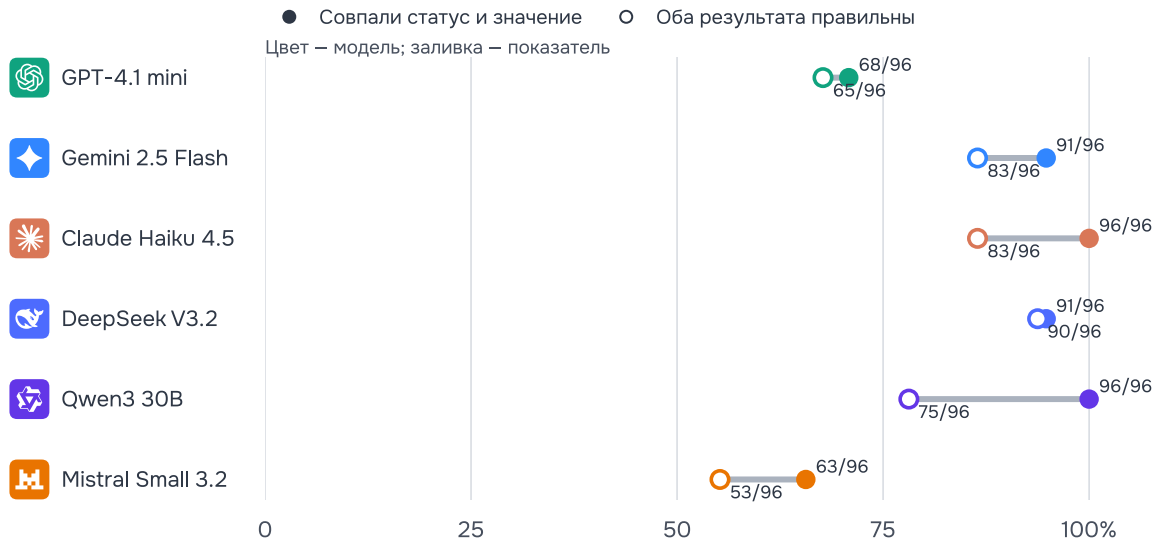
7. Качество эталонов и последующий анализ чувствительности

8. Воспроизводимость

Порядок цитирования

1. Предмет исследования

Рисунок 1. Повторяемость не гарантирует правильность



Дополнительный расчет после извлечения JSON, 96 пар на модель. Цвет и знак обозначают модель. Заполненная точка показывает совпадение статуса и значения в двух ответах, контурная — пары, в которых оба ответа совпали с эталоном. Проверка ссылок в этот показатель не входит. Недопустимые ответы и ошибки API остаются в полном парном знаменателе. Источник: архив воспроизводимости серии, файл summary-extracted.json.

Содержательная правильность, совпадение ответов и соблюдение формата рассматриваются раздельно: устойчивое воспроизведение одной ошибки требует иного решения, чем случайное нарушение структуры правильного ответа. У трех моделей исходная строгая обработка не приняла ни одного базового ответа, тогда как после заранее определенного извлечения единственного блока число правильных предметных ответов составило 83/96 у Haiku, 83/96 у Gemini и 60/96 у Mistral.

Работа — третья в серии CP-RU-DOC-2026-01, проведенной по Методике сравнительных испытаний языковых моделей на русском языке; ответы по документам и устойчивость к посторонним указаниям разобраны в двух других отчетах серии, общая постановка вопросов — в исследовательской повестке Центра.

2. Корпус и устройство сравнения

Основная серия — 96 синтетических ситуаций шести семейств и шесть моделей (методика, разделы 2 и 3); обращения выполнены через OpenRouter, исходные тексты и ответы раскрываются в общем архиве. Сведения о полученном поставщике следует читать вместе с идентификатором модели: общий интерфейс допускает разные маршруты исполнения.

Для каждой модели исходный вопрос предъявлен дважды отдельными обращениями. Третье обращение содержит заранее подготовленное перефразирование при тех же документах и эталоне. В запросе заданы температура ноль, предел 384 выходных токена и отключение дополнительного рассуждения. Фактическое применение параметров различается между сервисами; совпадение определяется по полученным конечным ответам. Один повтор показывает наблюдаемое поведение конкретной пары обращений; распределение ответов при многократном повторении требует отдельной серии.

Общий эксперимент включает восемь условий и 4608 запланированных ответов. В настоящем отчете основные сопоставления относятся к исходному вопросу, его повтору и перефразированию; остальные условия входят в общую ведомость расходов. Техническая серия на шести других заданиях отделена от основных результатов. Разделение важно для учета затрат: стоимость проверки интерфейса существует, но ее ответы не увеличивают объем основного корпуса.

3. Два уровня оценки формата

Дополнительный анализ, зафиксированный до основного запуска, допускает извлечение единственного блока JSON из разметки; правило — методика, раздел 5. Сопоставление двух оценок показывает, какая часть потери связана с представлением ответа, а какая сохраняется после извлечения объекта.

Модель	Схема: строго → извлечение	Правильно, строго; 95 % интервал	Правильно после извлечения; 95 % интервал	С доказательствами: строго → извлечение
GPT-4.1 mini	66/96 → 76/96	62/96 (54,6–73,4 %)	72/96 (65,5–82,6 %)	45/96 → 55/96
Gemini 2.5 Flash	0/96 → 91/96	0/96 (0,0–3,8 %)	83/96 (78,2–91,9 %)	0/96 → 52/96
Claude Haiku 4.5	0/96 → 96/96	0/96 (0,0–3,8 %)	83/96 (78,2–91,9 %)	0/96 → 68/96
DeepSeek V3.2	82/96 → 95/96	79/96 (73,5–88,6 %)	91/96 (88,4–97,8 %)	79/96 → 90/96
Qwen3 30B A3B	96/96 → 96/96	75/96 (68,9–85,2 %)	75/96 (68,9–85,2 %)	28/96 → 28/96
Mistral Small 3.2	0/96 → 74/96	0/96 (0,0–3,8 %)	60/96 (52,5–71,5 %)	0/96 → 35/96

После извлечения оценивается точное соответствие канонических полей эталону. Это закрытая проверка заданных статусов, значений и идентификаторов документов. Числа не заменяют содержательную экспертизу произволь-

ных свободных ответов. Для нулевых строгих результатов интервал доли правильных ответов составляет 0,0–3,8 %; причина полного провала формата раскрывается второй процедурой.

Диагностика базового оформления: у Haiku все 96 объектов извлечены из блока, в 88 ответах присутствует внешний текст; у Gemini блок имеется во всех 96 ответах, схему проходят 91. У Mistral обнаружены 93 разбираемых блока и три блока с неправильным JSON, восстановлены 74 допустимых объекта. У DeepSeek имеются 82 исходных JSON, 13 блоков и один отсутствующий для разбора ответ; восстановлены 13 объектов. У GPT-4.1 mini — 82 исходных JSON и 14 блоков, восстановлены 10; у Qwen все 96 ответов представлены исходным JSON. Извлечение объекта и прохождение схемы учитываются раздельно.

4. Совпадение, правильность и ссылки

Основное совпадение определяется по паре «статус, нормализованное значение». Внешние пробелы удаляются, регистр букв не учитывается, чисто числовые строки сопоставляются как десятичные числа. Единицы измерения, внутренние пробелы и пунктуация сохраняют значение для проверки. Самооценка уверенности в определение совпадения не входит: иначе изменение одного служебного числа скрывало бы воспроизведение предметного ответа.

Отдельное совпадение доказательств требует также равенства множеств ссылок. Порядок ссылок и их повторы значения не имеют. Один и тот же правильный ответ может сопровождаться разными допустимыми дополнительными ссылками; поэтому разница между двумя показателями имеет самостоятельный смысл. Полная доказательная правильность требует включения всех необходимых идентификаторов и отсутствия ссылок за пределами предоставленных документов. Это проверка соблюдения заранее заданного набора ссылок в закрытом корпусе.

Рядом с совпадением приводятся пары с двумя совпадениями с эталоном, пары с двумя одинаковыми расхождениями с эталоном и пары хотя бы с одним недопустимым ответом. Все доли используют полный парный знаменатель. Совпадение среди оставшихся после отбрасывания ошибок было бы условной характеристикой и скрывало бы часть сбоев. Повторение статуса или значения, расходящегося с эталоном, означает устойчивое расхождение. Для содержательного вывода об ошибке модели требуется также однозначность задания и эталона.

Перефразирование оценивается дополнительно через парное изменение правильности: результат по измененному вопросу минус результат по исходному. Положительная разница означает больше правильных решений на тех же случаях, отрицательная — меньше. Равенство общих долей допускает взаимную замену правильных и неправильных случаев. Поэтому нулевая разница читается вместе с парными совпадениями и исходными ответами. Во всех 16

арифметических перефразированиях выявлено нарушение падежного управления; смысл обычно понятен, но измеряемое изменение смешивает способ формулировки и ее грамматическое качество.

Для долей приводятся описательные 95-процентные интервалы Уилсона, для разниц — повторные выборки 24 шаблонных кластеров внутри семейств (методика, раздел 7).

Строгая обработка: повтор

Модель	Статус и значение; 95 % интервал	Также ссылки	Оба правильны	Оба расходятся с эталоном и совпали	Расходятся с эталоном, совпали со ссылками	Недопустимые пары
GPT-4.1 mini	57/96 (49,4–68,7 %)	54/96	55/96	2/96	2/96	39/96
Gemini 2.5 Flash	0/96 (0,0–3,8 %)	0/96	0/96	0/96	0/96	96/96
Claude Haiku 4.5	0/96 (0,0–3,8 %)	0/96	0/96	0/96	0/96	96/96
DeepSeek V3.2	74/96 (67,7–84,4 %)	63/96	73/96	1/96	1/96	20/96
Qwen3 30B A3B	96/96 (96,2–100,0 %)	93/96	75/96	21/96	20/96	0/96
Mistral Small 3.2	0/96 (0,0–3,8 %)	0/96	0/96	0/96	0/96	96/96

Извлечение единственного блока: повтор

Модель	Статус и значение; 95 % интервал	Также ссылки	Оба правильны	Оба расходятся с эталоном и совпали	Расходятся с эталоном, совпали со ссылками	Недопустимые пары
GPT-4.1 mini	68/96 (61,1–79,0 %)	64/96	65/96	3/96	3/96	28/96
Gemini 2.5 Flash	91/96 (88,4–97,8 %)	90/96	83/96	8/96	8/96	5/96
Claude Haiku 4.5	96/96 (96,2–100,0 %)	96/96	83/96	13/96	13/96	0/96

Модель	Статус и значение; 95 % интервал	Также ссылки	Оба правильны	Оба расходятся с эталоном и совпали	Расходятся с эталоном, совпали со ссылками	Недопустимые пары
DeepSeek V3.2	91/96 (88,4–97,8 %)	76/96	90/96	1/96	1/96	1/96
Qwen3 30B A3B	96/96 (96,2–100,0 %)	93/96	75/96	21/96	20/96	0/96
Mistral Small 3.2	63/96 (55,7–74,4 %)	62/96	53/96	10/96	10/96	30/96

У Qwen полное повторение предметного ответа сочетается с 75 парами, совпадающими с эталоном, и 21 парой с одинаковым расхождением с эталоном. Совпадение вместе со ссылками наблюдалось в 93/96 пар. У Haiku после извлечения совпали все 96 предметных ответов и все наборы ссылок, но 83 пары совпали с эталоном, а 13 расходились с ним. У DeepSeek после извлечения предметное совпадение составило 91/96, а совпадение вместе со ссылками — 76/96: стабильность значения и стабильность доказательного оформления различаются.

Перефразирование: парное изменение предметной правильности

Модель	Строго: разница, п.п. [95 % интервал]	После извлечения: разница, п.п. [95 % интервал]
GPT-4.1 mini	-1,04 [-9,38; +8,33]	+1,04 [-5,21; +7,29]
Gemini 2.5 Flash	+0,00 [+0,00; +0,00]	-2,08 [-6,25; +2,08]
Claude Haiku 4.5	+0,00 [+0,00; +0,00]	+1,04 [+0,00; +3,12]
DeepSeek V3.2	+3,12 [-3,12; +10,42]	+0,00 [-4,17; +4,17]
Qwen3 30B A3B	-1,04 [-5,21; +3,12]	-1,04 [-5,21; +3,12]
Mistral Small 3.2	+0,00 [+0,00; +0,00]	+8,33 [-5,21; +22,92]

Каждая разница рассчитана по 96 парам. Изменение у Mistral после извлечения составляет +8,33 процентного пункта, но кластерный интервал простирается от −5,21 до +22,92 пункта. Это наблюдаемое изменение данной выборки

с существенной неопределенностью. Нулевые строгие разницы у Haiku, Gemini и Mistral возникают при полном отклонении формата в обоих условиях и не устанавливают устойчивость предметного решения.

5. Уверенность и передача на другую модель

Число confidence трактуется как самооценка модели относительно правильности полного предметного ответа, включая статус. Оценка проводится только для объектов, прошедших схему: ошибки формата раскрываются отдельно. Показатель Брайера усредняет квадрат отклонения самооценки от фактического исхода, а ошибка калибровки сопоставляет среднюю уверенность и долю правильных решений в пяти заранее заданных интервалах.

Для применения существенна численность интервалов. Малое число ответов при низкой уверенности оставляет широкий диапазон возможного поведения вне этой серии. Особенно важны ошибки с высокой уверенностью: правило передачи только низкоуверенных случаев сохраняет такие ответы в первом звене. Правильный средний показатель также может скрывать отдельное семейство, в котором уверенность заметно превышает правильность.

Дополнительный расчет использует фактически полученные исходные ответы двух моделей. Сначала берется менее затратная модель по средней фактической стоимости обращения. Если ее ответ нарушает схему или уверенность ниже 80, вместо него выбирается ответ более затратной модели на том же задании. Порог зафиксирован до основной серии. Эталонная правильность в условие передачи не входит; истинный исход используется лишь для последующей оценки выбранного ответа.

Стоимость маршрута включает первый вызов всегда и второй при передаче. Суммирование только цены выбранного ответа занизило бы расходы. Сравнение выполняется при полном парном покрытии всех 96 случаев и известных затратах обеих моделей. Это ретроспективное воспроизведение правила на сохраненных ответах. Для дополнительного извлечения из блока расчет повторяется отдельно, поскольку исправление формата меняет условие передачи уже на первом шаге.

Содержательная интерпретация требует проверить три числа одновременно: долю передач, правильность итогового выбора и общую стоимость. Если почти все ответы первого звена проходят формат и содержат высокую уверенность, второе звено вызывается редко даже при предметных ошибках. Если первый ответ часто отклоняется по форме, маршрут оплачивает оба обращения. Таким образом, дешевый первый вызов создает экономию только при определенном распределении уверенности, ошибок и затрат; это распределение измеряется по конкретным парам.

Модель	Допустимых: строго / извлечение	Брайер: строго / извлечение	Ошибка калибровки: строго / извлечение
GPT-4.1 mini	66 / 76	0.0714 / 0.0620	0.0379 / 0.0329
Gemini 2.5 Flash	0 / 91	нет оценки / 0.0879	нет оценки / 0.0879
Claude Haiku 4.5	0 / 96	нет оценки / 0.2015	нет оценки / 0.1448
DeepSeek V3.2	82 / 95	0.0336 / 0.0405	0.0174 / 0.0198
Qwen3 30B A3B	96 / 96	0.3111 / 0.3111	0.3260 / 0.3260
Mistral Small 3.2	0 / 74	нет оценки / 0.3480	нет оценки / 0.3581

После извлечения у Qwen 77 допустимых ответов попали в интервал уверенности 80–100 %, но 19 из 77 расходились с предметным эталоном; средняя уверенность составила 99,7 %, правильность — 75,3 %. У Gemini все 91 допустимый ответ имел уверенность 100 %, среди них восемь расхождений с эталоном. У DeepSeek в верхнем интервале было 94 ответа и четыре расхождения с эталоном; у GPT-4.1 mini — 69 и три; у Haiku — 64 и семь; у Mistral — 38 и пять. Эти условные числа описывают сохранение высокоуверенных расхождений с фиксированным эталоном первым звеном.

Расчет маршрута Mistral → Haiku

Для подробного разбора выбрана самая дешевая и самая дорогая базовые модели среди пар с полным учетом расходов. Это правило выбора использует наблюдаемую стоимость, а не итоговую правильность. Из 30 направленных пар разных моделей 10 исключены из-за неполного учета Gemini и еще 10 — из-за отсутствия более дешевого первого звена; 10 допустимых маршрутов раскрыты в архиве. Все результаты маршрута рассчитаны на 96 случаях.

При строгом приеме все 96 ответов Mistral требуют передачи, но Haiku также не проходит формат: правильность остается 0/96. Общая цена равна $0,0033289 + $0,173115 = $0,1764439. При извлечении блока передаются 58/96 случаев (60,4 %; интервал 50,4–69,6 %). Первый этап стоит $0,0033289, реально запрошенные ответы Haiku на выбранных 58 случаях — $0,104649; итог $0,1079779. Стоимость одного случая равна $0,00112477.

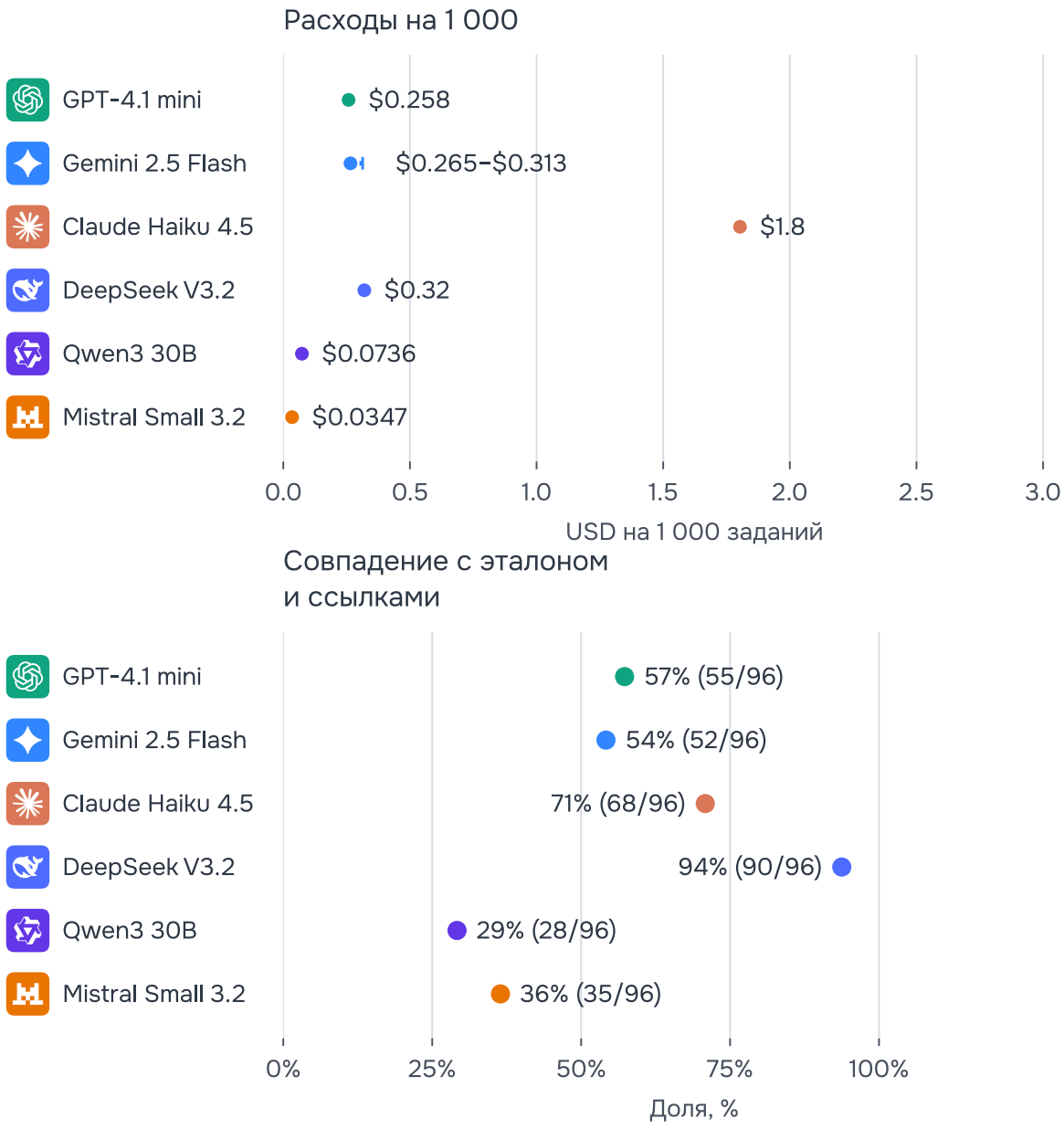
Вторичный маршрут дает 87/96 правильных предметных результатов (90,6 %; интервал 83,1–95,0 %) и 73/96 доказательно правильных (76,0 %; 66,6–83,5 %). Соответствующие цены — $0,1079779 / 87 = $0,00124113 и $0,1079779 / 73 = $0,00147915. При прямом вызове Haiku после извлечения получено 83/96 и 68/96 за $0,173115. Сопоставление описывает эти сохраненные пары и фиксированный порог; подбора порога и проверки на отложенной выборке в серии нет.

Для сопоставления маршрута с другими доступными процедурами существенен прямой вызов DeepSeek: после извлечения блока он дает 91/96 совпадений предметных полей с эталоном и 90/96 результатов с требуемыми ссылками

за $0,0307115592. На исходном корпусе эти числа выше, а зарегистрированные расходы ниже, чем у маршрута Mistral → Haiku с 87/96 и 73/96 за $0,1079779. Поэтому сопоставление только с дорогим прямым вызовом Haiku недостаточно для выбора процедуры. Это наблюдение относится к данным интерфейсам, различающемуся фактическому выполнению и ограниченному корпусу; оно не устанавливает универсального преимущества модели.

6. Учет расходов и времени

Рисунок 2. Цена обращения и результат по эталону



Расходы пересчитаны на 1 000 заданий из наблюдаемой базовой серии. Для Gemini показан диапазон учета от подтвержденной суммы до суммы с резервом, а не доверительный интервал. Нижняя панель относится к исходным 96 заданиям после извлечения JSON. Источник: архив воспроизводимости серии, файлы accounting.json и summary-extracted.json.

Фактическая сумма берется из сведений об использовании интерфейса. В каждую ячейку входят все технические попытки, а правильность определяется по последней попытке. Входные, выходные и общие токены учитываются отдельно; поля кеша сохранены в исходных ответах. Отсутствующее поле сохраняется как отсутствие сведений.

Стоимость одного обращения равна сумме расходов ячейки, деленной на число заданий. Стоимость правильного предметного результата делит ту же сумму на число правильных статусов и значений. Стоимость доказательно правильного результата использует более строгий знаменатель — правильный предметный ответ со всеми необходимыми и допустимыми ссылками. При нуле правильных результатов конечная стоимость успешного случая не определена; одна низкая цена запроса не разрешает этот исход.

Сравнение использует зарегистрированный расход; фактическая сумма зависит от выхода, кеширования и поставщика. Тарифы сохранены в каталоге на дату запуска. Доллары США — единица счета интерфейса; пересчет в рубли потребовал бы учета курса и условий платежа.

Задержка измерена на стороне клиента от начала обращения до получения результата. Для повторных попыток учитывается сумма их длительностей. Медиана и 95-й процентиль описывают этот запуск, включая сеть, очередь, посредника и генерацию. Сумма времени двух сохраненных вызовов при передаче — расчетная длительность последовательного маршрута. Производственная доступность, предельная пропускная способность и договорный уровень обслуживания требуют другого плана измерений.

Модель	USD / случай, 96 случаев	USD / правильный, строго	USD / правильный после извлечения	USD / доказательно правильный после извлечения
GPT-4.1 mini	$0.00025774	$0.00039908	$0.00034365	$0.00044987
Gemini 2.5 Flash	$0.00026532 → $0.00031333	нет конечной оценки	$0.00030687 → $0.00036240	$0.00048982 → $0.00057845
Claude Haiku 4.5	$0.00180328	нет конечной оценки	$0.00208572	$0.00254581
DeepSeek V3.2	$0.00031991	$0.00038875	$0.00033749	$0.00034124
Qwen3 30B A3B	$0.00007355	$0.00009415	$0.00009415	$0.00025219
Mistral Small 3.2	$0.00003468	нет конечной оценки	$0.00005548	$0.00009511

Диапазоны Gemini обозначают подтвержденный подытог → консервативный бюджетный учет, а не статистический интервал или гарантированную границу счета поставщика. Подтвержденный базовый подытог всех 96 успешных ответов — $0,0254704, бюджетный учет с двумя попытками без сведений

об оплате — $0,0300796. Деление только полного поднабора из 94 заданий потеряло бы известные расходы двух остальных. Конечная полная цена Gemini остается неизвестной.

Для остальных моделей подтвержденные базовые суммы составляют $0,0247428 у GPT-4.1 mini, $0,173115 у Haiku, $0,0307115592 у DeepSeek, $0,0070612 у Qwen и $0,0033289 у Mistral. У Qwen одна и та же сумма дает $0,0070612 / 75 = $0,00009415 за предметно правильный ответ и $0,0070612 / 28 = $0,00025219 за доказательно правильный. Поэтому сравнение одной цены вызова скрывает различие конечного результата.

Модель	Входные / выходные токены базовой серии	Медиана / 95-й процентиль времени, с
GPT-4.1 mini	47021 / 3709 (96 случаев)	1.69 / 2.22
Gemini 2.5 Flash	46017 / 4454 (94 полных случая)	1.34 / 1.80
Claude Haiku 4.5	68560 / 20911 (96 случаев)	2.75 / 5.21
DeepSeek V3.2	52609 / 63596 (96 случаев)	7.71 / 35.90
Qwen3 30B A3B	58891 / 3907 (96 случаев)	1.66 / 2.63
Mistral Small 3.2	56587 / 4103 (96 случаев)	1.62 / 3.14

Временные показатели рассчитаны для всех 96 базовых заданий каждой модели; они суммируют длительности попыток данного задания, без ожидания между повторами и очереди планировщика.

DeepSeek сообщил токены рассуждения вопреки переданному отключению этого режима (цифры аудита — методика, раздел 3). Эти данные характеризуют отчетность интерфейса: одинаковые запрошенные параметры не устанавливают равенство фактического вычислительного бюджета. Цена и задержка отражают наблюдаемое выполнение сервисов, поэтому чистый рейтинг внутреннего качества моделей из такого сравнения не следует.

Во всей основной матрице завершены 4608 заданий; выполнено 4649 попыток, включая 41 повторную у 27 заданий. Зарегистрировано 46 ошибок попыток: 45 ответов HTTP 429 и один разрыв соединения; пять заданий завершились ошибкой, все в условиях с посторонними указаниями. В базовой серии конечных ошибок нет. Подтвержденные расходы основной серии — $2,35782682580; технической серии из 216 вызовов — $0,07611246625; вместе — $2,43393929205. Для 46 попыток сумма оплаты отсутствует. Общий бюджетный учет $2,54614139205 включает резервные подстановки и отличается от подтвержденного счета.

Сведения поставщиков основной серии: OpenAI для GPT-4.1 mini; Google AI Studio для успешных ответов Gemini; Amazon Bedrock и один ответ Azure для Haiku; GMICloud и DeepInfra для DeepSeek; Nebius, SiliconFlow и DekaLLM для Qwen; Mistral, Venice и DeepInfra для Mistral Small. Это распределение всей матрицы, а не причинное объяснение качества. Сведения кеша доступны в исходных ответах; отдельная оценка эффекта кеширования в план сравнения не включена.

7. Качество эталонов и последующий анализ чувствительности

Повторный аудит после получения ответов выявил неоднозначность шаблона arithmetic-t3, где не заданы цена и денежная единица, и более слабое допущение в arithmetic-t4, где не задан начальный запас (описание дефекта — методика, раздел 9).

Основные таблицы на 96 случаях сохранены как запись зафиксированного эксперимента. Дополнительно, после основного запуска, выполнены два анализа чувствительности: исключение целиком arithmetic-t3 оставляет 92 случая и 23 шаблонных кластера; исключение также arithmetic-t4 — 88 случаев и 22 кластера.

Корпус	Обработка	Повтор Qwen	Совпавшие расхождения Qwen; 95 % интервал	Передачи Mistral → Haiku	Маршрут: предметно	Маршрут: со ссылками	Цена маршрута
Исходный корпус, 96	Строго	96/96	21/96 (14,8–31,1 %)	96/96	0/96	0/96	$0.17644 390000
Исходный корпус, 96	Извлечение	96/96	21/96 (14,8–31,1 %)	58/96	87/96	73/96	$0.107977 90000
Без неоднозначной цены, 92	Строго	92/92	19/92 (13,6–30,0 %)	92/92	0/92	0/92	$0.168910 10875
Без неоднозначной цены, 92	Извлечение	92/92	19/92 (13,6–30,0 %)	54/92	86/92	72/92	$0.10044 410875
Без цены и начального остатка, 88	Строго	88/88	19/88 (14,3–31,3 %)	88/88	0/88	0/88	$0.16067 902875

Корпус	Обработка	Повтор Qwen	Совпавшие расхождения Qwen; 95 % интервал	Передачи Mistral → Haiku	Маршрут: предметно	Маршрут: со ссылками	Цена маршрута
Без цены и начального остатка, 88	Извлечение	88/88	19/88 (14,3–31,3 %)	52/88	82/88	68/88	$0.09619202875

В таблице предметная и доказательная правильность означает соответствие сохраненным эталонам и наборам ссылок. Совпадающие расхождения Qwen уменьшаются с 21/96 до 19/92; дополнительное исключение начального остатка оставляет 19/88. Таким образом, полное воспроизведение не совпадающих с эталоном результатов сохраняется вне двух обсуждаемых конструкций. При этом исходные 21 случая нельзя целиком описывать как однозначные ошибки модели. У Qwen среди допустимых базовых ответов с уверенностью от 80 % сохраняются 19 расхождений с эталоном из 76 на сокращенном корпусе из 92 заданий и 19 из 73 на корпусе из 88 заданий.

При извлечении блока маршрут сохраняет предметный результат 86/92 и 82/88 при 54 и 52 передачах соответственно; строгая процедура продолжает отклонять ответы обоих звеньев. Снижение общей цены в сокращенных строках связано также с меньшим числом заданий. Полные пересчеты, интервалы и исключенные идентификаторы включены в reference-quality.json архива воспроизводимости.

8. Воспроизводимость

Архив воспроизводимости содержит корпус, генератор, зафиксированные планы, параметры, каталог моделей, конечное содержимое ответов, сведения usage, учет попыток, агрегаты обеих процедур и документы аудита. Формат учета usage и сведений маршрутизации описан в документации OpenRouter; фактическим основанием опубликованных чисел служат сохраненные ответы данной серии.

Порядок цитирования

Федеральный центр экспертизы в сфере безопасности и развития искусственного интеллекта имени Д. А. Поспелова. Повторяемость ответов языковых моделей и стоимость правильного результата: исследование CP-RU-DOC-2026-01. 11 ноября 2026 года. При цитировании чисел следует указывать процедуру обработки — строгий прием либо извлечение единственного блока — и сохранять знаменатель. Замечания по материалам направляются на expertise@pospelov.center.