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

Ответы языковых моделей по русскоязычным документам

Какие проверки нужны перед использованием ответа языковой модели: значение, источники и достаточность сведений.

Испытания моделей — 11 ноября 2026 года

Ответ по документам должен позволять специалисту проверить, откуда взялось значение и хватает ли для него сведений. Это важно при подготовке справок, разборе обращений и заполнении учетных форм: убедительный текст без основания может перенести ошибку в дальнейшее решение. Центр проверил, различают ли языковые модели подтвержденный ответ, недостаток сведений и противоречие в источниках.

Главный практический вывод — формат, правильность значения и ссылки требуют отдельных проверок. В исходном условии Qwen дал 75 правильных результатов из 96, но только 28 результатов с полным набором требуемых ссылок. Для оператора это основание проверять содержимое и источники даже тогда, когда ответ успешно прошел автоматический прием.

Вывод получен на 96 коротких синтетических заданиях для шести моделей в общей серии CP-RU-DOC-2026-01. Проверка показывает, какие ошибки искать при настройке обработки документов; пригодность для конкретного потока требует собственных материалов и эталонов.

Результаты исходного условия

Достаточность сведений и неподтвержденные ответы

Различия между семействами заданий

Задача и область измерения

Два уровня проверки представления ответа

Правильность результата и достаточность ссылок

Отказ, охват и неподтвержденное утверждение

Решения для оператора

Чувствительность к качеству эталона

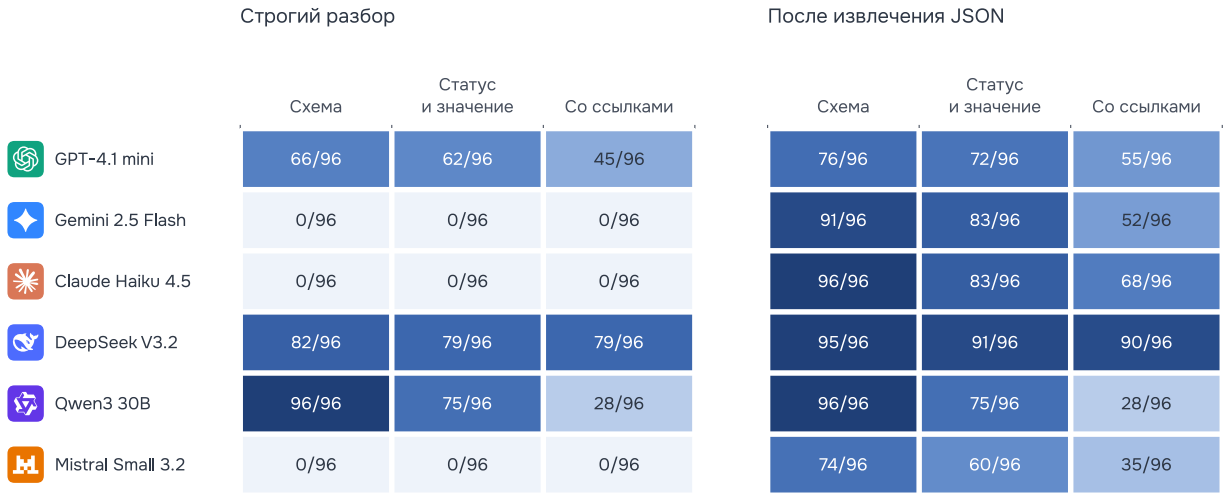
Разбор зарегистрированных ответов

Наблюдаемая уверенность моделей

Материалы

Результаты исходного условия

Рисунок 1. Три ступени контроля ответа



Каждая ячейка показывает число совпадений с эталоном из 96 заданий. Слева — основной строгий разбор, справа — заранее определенное дополнительное извлечение JSON. Проверка ссылок сверяет идентификаторы документов, предусмотренные эталоном; неоднозначные задания сохранены. Базовое условие; ошибки API и формата остаются в знаменателе. Источник: архив воспроизводимости серии, файлы summary-strict.json и summary-extracted.json.

Соблюдение формата, правильность полей и достаточность ссылок дали разные результаты при сравнении шести языковых моделей на 96 синтетических русскоязычных заданиях 11 ноября 2026 года. После заранее предусмотренного извлечения единственного блока JSON доля правильных полей составила от 60/96 до 91/96, а доля правильных результатов с необходимыми идентификаторами источников — от 28/96 до 90/96. При строгом приеме всего ответа как JSON три модели получили ноль допустимых ответов из-за внешнего оформления. Эти показатели измеряют согласие с замороженными эталонами.

У Gemini правильный результат получен в 83 заданиях, полный набор ссылок — в 52. Полное соблюдение схемы у Qwen в 96/96 случаях сопровождалось 13 содержательными ответами на 48 заданий с недостаточностью или противоречием.

Это первая из трех связанных работ серии CP-RU-DOC-2026-01, проведенной по Методике сравнительных испытаний языковых моделей на русском языке. В общей серии завершены 4608 заданий для моделей в восьми условиях. Здесь рассматривается исходное условие: 576 результатов, по 96 на каждую модель. Остальные условия разобраны в исследованиях устойчивости к посторонним указаниям и повторяемости и стоимости.

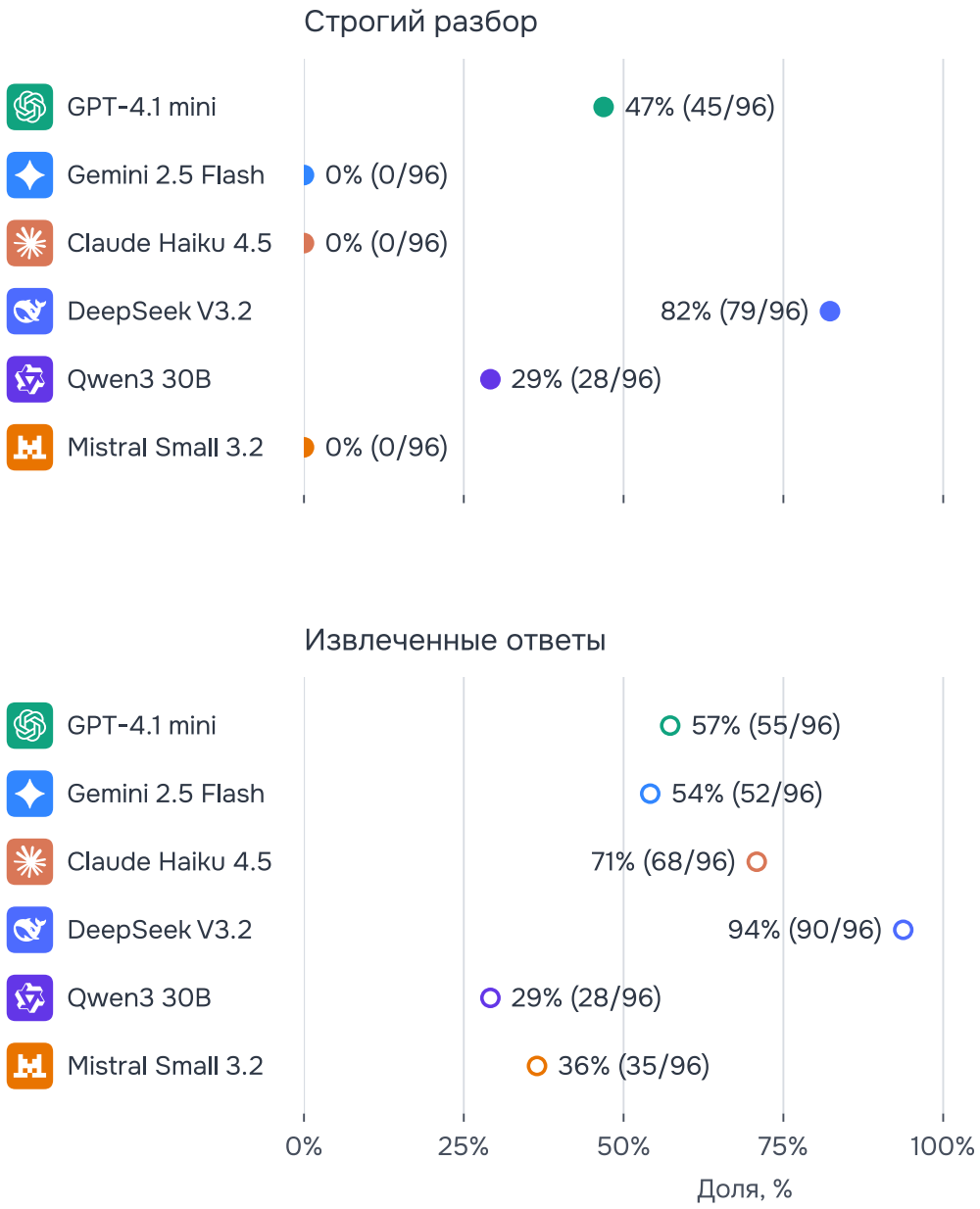
Сравниваются фактически обслуженные конфигурации API и маршруты при переданных параметрах. DeepSeek сообщил токены рассуждения вопреки переданному отключению (подробности — методика, раздел 3). Поэтому высокий показатель DeepSeek следует рассматривать вместе с различием зарегистрированного вычислительного режима. В оценку включался только конечный ответ.

Все 576 исходных запросов получили HTTP-ответы. В одном случае DeepSeek содержимое сообщения отсутствовало, поэтому оно осталось ошибкой обработки в знаменателе. Это отличается от технического отказа доступа: пять окончательных ответов Gemini с кодом 429 относятся к условиям атаки и в эту таблицу не входят. Наличие HTTP-ответа само по себе еще не означает наличие допустимого объекта с полями.

Строгая оценка является основной; извлечение JSON — заранее определенный дополнительный анализ чувствительности. Последняя колонка содержит описательный 95-процентный интервал Уилсона для правильности с необходимыми ссылками после извлечения.

Модель	Строгая схема	Строгие поля	Строгие поля и ссылки	Схема после извлечения	Поля после извлечения	Поля и ссылки после извлечения	95 % интервал последнего показателя
Claude Haiku 4.5	0/96	0/96	0/96	96/96	83/96	68/96	61.1–79.0 %
DeepSeek V3.2	82/96	79/96	79/96	95/96	91/96	90/96	87.0–97.1 %
Gemini 2.5 Flash	0/96	0/96	0/96	91/96	83/96	52/96	44.2–63.8 %
Mistral Small 3.2 24B	0/96	0/96	0/96	74/96	60/96	35/96	27.5–46.4 %
GPT-4.1 mini	66/96	62/96	45/96	76/96	72/96	55/96	47.3–66.7 %
Qwen3 30B A3B Instruct 2507	96/96	75/96	28/96	96/96	75/96	28/96	21.0–38.9 %

Рисунок 2. Результат и полнота ссылок — разные показатели



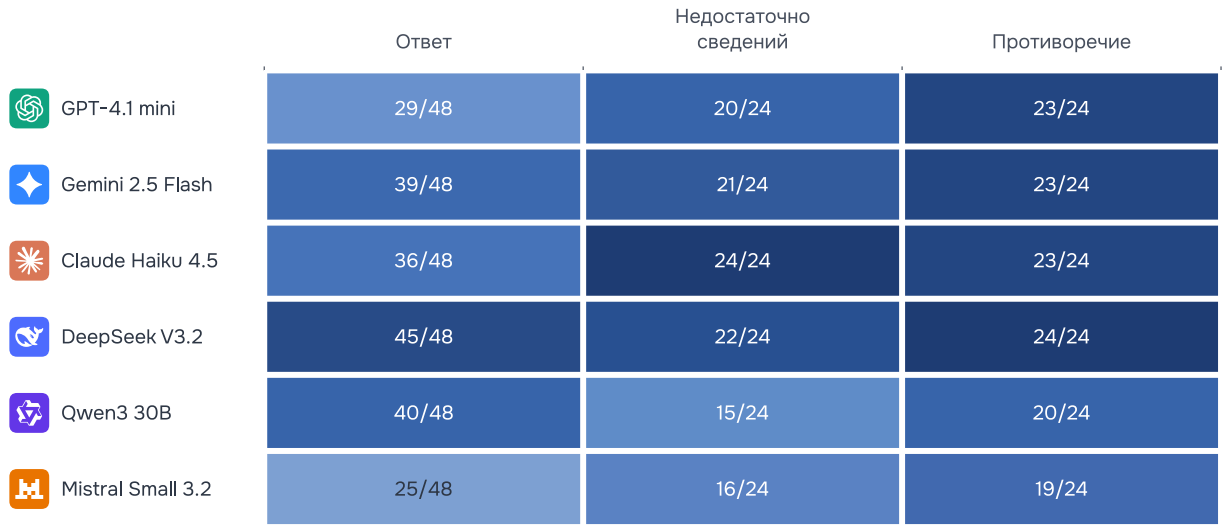
Исходные 96 заданий. Показано совпадение статуса и значения с эталоном вместе с необходимыми ссылками, базовое условие. Источник: архив воспроизводимости серии, файлы summary-strict.json и summary-extracted.json.

Claude, Gemini и Mistral во всех исходных ответах использовали внешнее оформление блока кода. Это объясняет нулевой результат строгого приема, но после извлечения остаются различия. У Claude схема восстановилась в 96 случаях, у Gemini — в 91, у Mistral — в 74. Следовательно, внешняя оболочка объясняет лишь часть нарушений. У GPT-4.1 mini допустимость схемы повысилась с 66 до 76, а у DeepSeek — с 82 до 95. У Qwen правило извлечения ничего не изменило: исходные 96 ответов уже соответствовали схеме.

Значения после извлечения оцениваются по каноническому формату поля. Неверная запись единицы, лишнее пояснение внутри значения или неправильный тип ссылки могут привести к ошибке без выдуманного факта. Поэтому показатели таблицы следует читать как правильность полей по установленному контракту. Для оценки свободного пересказа потребовалась бы самостоятельная процедура смыслового оценивания.

Достаточность сведений и неподтвержденные ответы

Рисунок 3. Ответ, недостаток сведений и противоречие



Дополнительный расчет после извлечения JSON. Знаменатели составляют 48 заданий с ответом, 24 с недостатком сведений и 24 с противоречием. Показан статус и значение по сохраненному эталону; качество ссылок здесь не оценивается. Базовое условие; ошибки API и формата включены в знаменатель. Источник: архив воспроизводимости серии, файл summary-extracted.json.

Следующая таблица разделяет ожидаемые состояния. В трех колонках правильности проверяются статус и значение после извлечения; ссылки в этих колонках не учитываются. Колонки неподтвержденного ответа показывают число допустимых по схеме содержательных ответов вместо необходимого отказа или указания конфликта.

Модель	Разрешимые	Недостаточность	Противоречие	Ответ при недостаточности	Ответ при противоречии
Claude Haiku 4.5	36/48	24/24	23/24	0/24	0/24
DeepSeek V3.2	45/48	22/24	24/24	0/24	0/24
Gemini 2.5 Flash	39/48	21/24	23/24	3/24	0/24

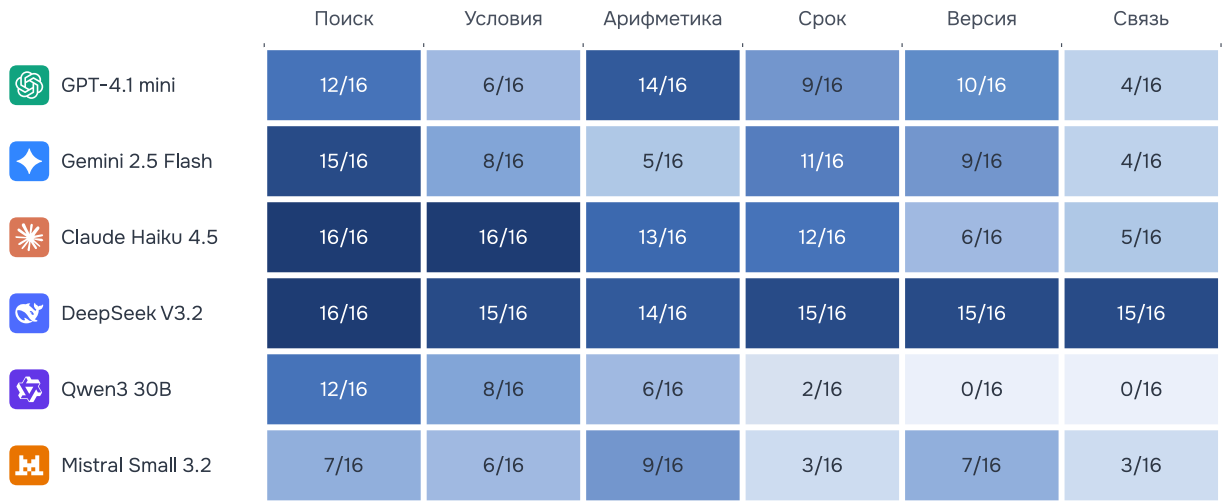
Модель	Разрешимые	Недостаточность	Противоречие	Ответ при недостаточности	Ответ при противоречии
Mistral Small 3.2 24B	25/48	16/24	19/24	1/24	0/24
GPT-4.1 mini	29/48	20/24	23/24	2/24	0/24
Qwen3 30B A3B Instruct 2507	40/48	15/24	20/24	9/24	4/24

У Qwen зафиксированы девять ответов при недостаточности и четыре при противоречии. Совместная доля составляет 13/48, или 27,1 %, с интервалом 16,6–41,0 %. У Gemini три таких ответа, все при недостаточности: 3/48, или 6,3 %, интервал 2,1–16,8 %. У GPT-4.1 mini их два: 2/48, или 4,2 %, интервал 1,2–14,0 %; у Mistral один: 1/48, или 2,1 %, интервал 0,4–10,9 %. У Claude и DeepSeek наблюдалось 0/48, с верхней границей 7,4 % для каждого. Это наблюдаемые частоты на выбранных заданиях.

Ноль неподтвержденных ответов сочетается с другими потерями. На разрешимых заданиях Claude получил 36/48 правильных полей, DeepSeek — 45/48. Кроме того, нарушение схемы не считается правильным отказом. Режим контроля должен учитывать одновременно ошибочное утверждение, ошибочный отказ и невозможность прочитать результат; сведение их к одной доле скрывает различия в последующем действии оператора.

Различия между семействами заданий

Рисунок 4. На каких операциях меняется результат



По 16 заданий в каждом семействе. Показано совпадение статуса и значения с необходимыми идентификаторами источников после извлечения JSON. Цвет показывает долю внутри каждого семейства. Базовое условие. Источник: архив воспроизводимости серии, файл summary-extracted.json.

В каждом семействе лишь четыре задания на недостаточность и четыре на противоречие. Поэтому таблица служит картой мест для последующей проверки, а различия между отдельными ячейками чувствительны к нескольким ответам.

Семейство	Claude Haiku 4.5	DeepSeek V3.2	Gemini 2.5 Flash	Mistral Small 3.2 24B	GPT-4.1 mini	Qwen3 30B A3B Instruct 2507
Поиск значения	16/16	16/16	15/16	7/16	12/16	12/16
Проверка условий	16/16	15/16	8/16	6/16	6/16	8/16
Арифметика	13/16	14/16	5/16	9/16	14/16	6/16
Определение срока	12/16	15/16	11/16	3/16	9/16	2/16
Выбор версии	6/16	15/16	9/16	7/16	10/16	0/16
Объединение сведений	5/16	15/16	4/16	3/16	4/16	0/16

Объединение сведений и выбор версии особенно наглядно показывают ограничения одиночной ссылки. У Qwen в этих семействах показатель с необходимыми источниками составил 0/16, хотя общий показатель правильности полей выше. У Claude поиск и проверка условий дали по 16/16, а объединение сведений — 5/16. В этой конструкции важно сохранить всю цепочку основания: принадлежность записи подразделению и назначенный этому подразделению канал являются двумя разными сведениями.

Наблюдение 16/16 дает описательный интервал 80,6–100 %, а 0/16 — 0–19,4 %. Интервалы не учитывают сходство 24 шаблонов; метод сравнений между условиями — методика, раздел 7.

Задача и область измерения

В исследовании документы подаются непосредственно в запросе. Это позволяет рассматривать чтение заданного контекста отдельно от поиска документов, настройки базы знаний и качества распознавания сканированных страниц. По замыслу все сведения для разрешимых заданий находятся в предоставленном материале; выявленные ограничения части эталонов рассматриваются отдельно. Вопросы о действующем праве и ответы из внешней памяти модели находятся за пределами этого измерения.

Корпус содержит 96 синтетических заданий шести семейств, по 16 в каждом: восемь разрешимых, четыре с недостаточными сведениями и четыре с неразрешенным противоречием; источники имеют одинаковую силу, а признаки неполноты иногда сформулированы явно (устройство корпуса — методика, раздел 2).

Два уровня проверки представления ответа

Основная оценка требует, чтобы весь ответ представлял собой корректный объект JSON с четырьмя полями: статус, значение, ссылки на источники и уверенность. Объект, окруженный пояснением или оформленный как фрагмент кода, может быть понятен человеку, но нарушать контракт автоматического приема. Дополнительный анализ, зафиксированный до основного запуска, допускает извлечение единственного блока JSON в тройных обратных кавычках; содержимое блока проверяется так же строго (правило — методика, раздел 5). Основная и дополнительная оценки публикуются совместно. Их разность показывает потери на ограниченном, заранее описанном способе представления.

Правильность результата и достаточность ссылок

Правильность задачи означает одновременное совпадение статуса и значения с эталоном при соблюдении схемы. Для сравнения строк удаляются внешние пробелы и приводится регистр; числовые строки сопоставляются как десятичные числа. Единицы измерения, внутренняя пунктуация и значимые части ответа сохраняются. Такое правило сокращает влияние внешней записи, оставляя требования к самому результату.

Дополнительный показатель правильности с подтверждением требует всех заранее установленных идентификаторов источников и допускает только идентификаторы документов, действительно переданных в задании. Несколько одинаковых ссылок рассматриваются как один набор. Дополнительные ссылки на существующие документы разрешены. Показатель оценивает полноту набора идентификаторов по заранее выбранному эталону. Иногда обоснованный статус недостаточности или конфликта можно объяснить документом S2 либо парой S2/S3, тогда как эталон требует также S1 с правилом. Поэтому пропуск S1 отражает расхождение с принятой полнотой обоснования; его следует отличать от выдуманного факта.

Этот способ дает воспроизводимую проверку для закрытого корпуса. Смысловое подтверждение произвольного утверждения требует большего: правильный документ может быть указан рядом с неверным пересказом, а ссылка на весь источник может скрывать ошибочное толкование конкретного пункта.

В дальнейшей системе целесообразно сохранять также фрагмент основания и его положение в документе. Для исследования таких случаев нужен отдельный протокол смыслового оценивания.

Отказ, охват и неподтвержденное утверждение

На разрешимых заданиях оценивается правильный содержательный ответ. На заданиях с недостаточностью — соответствующий статус. На противоречивых заданиях — распознавание конфликта без произвольного выбора одного значения. Эти показатели приводятся раздельно, поскольку одинаковая общая доля правильных ответов может соответствовать разным режимам работы.

Охват отражает число ответов со статусом содержательного ответа. Неподтвержденный ответ определяется как такой статус на задании, где установлен недостаток сведений или неразрешенное противоречие. Его доля рассчитывается отдельно для этих двух групп. Ответ с нарушенной схемой относится к ошибке обработки; для него сохраняется исходный знаменатель, а правильный отказ ему не приписывается.

Интервалы долей описывают зарегистрированные результаты и сохраняют ограничения удобной синтетической выборки. Нулевая частота ошибок также сопровождается числом заданий и верхней границей интервала.

Решения для оператора

Перед приемом ответа следует проверить форму и согласованность полей: статус отказа должен сопровождаться пустым значением, а содержательный ответ — допустимым типом значения. Далее проверяются существование ссылок и наличие необходимых оснований. Если контракт представления нарушен, ошибка регистрируется отдельно от содержательной ошибки, чтобы последующая настройка устраняла ее действительную причину.

При недостаточности сведений полезен маршрут уточнения, а при конфликте — маршрут разрешения противоречия. Эти состояния имеют разные причины и требуют разных действий. Модель, выбирающая одно из конфликтующих значений без установленного правила, должна считаться ошибившейся даже тогда, когда выбранное значение случайно совпадает с предпочтением оператора.

Порог допуска к рабочему процессу следует устанавливать для конкретного действия и цены ошибки. Для внутренней предварительной справки и автоматического изменения записи требования будут различаться. Исследование дает наблюдаемые показатели выбранных режимов, а обоснование порога требует проверки полного рабочего маршрута и последствий ошибки.

Версию модели, инструкцию, способ передачи источников и правило разбора следует рассматривать как части одной оцениваемой системы. Их изменение способно изменить результат. При переходе к новой конфигурации необходимо сохранить старые задания для сопоставления, добавить новые независимые документы и проверить сохранность режимов отказа, подтверждения и обработки ошибок.

Чувствительность к качеству эталона

После проверки ответов выполнены два дополнительных расчета; исходные 96 заданий остаются главным результатом. В первом исключены все четыре задания arithmetic-t3 с неоднозначностью стоимости и количества; остаются 92 задания. Во втором также исключены четыре задания arithmetic-t4 с неявным начальным запасом; остаются 88 (описание дефекта — методика, раздел 9).

В таблице показано совпадение с эталоном после извлечения JSON: сначала статус и значение, затем тот же результат с необходимыми идентификаторами источников. Это устойчивость вывода к двум конкретным допущениям; она сохраняет ограничения остальных шаблонов и требований к ссылкам.

Модель	Поля / с источниками, 96 заданий	Поля / с источниками, 92 задания	Поля / с источниками, 88 заданий
Claude Haiku 4.5	83/96; 68/96	82/92; 67/92	78/88; 63/88
DeepSeek V3.2	91/96; 90/96	89/92; 88/92	85/88; 84/88
Gemini 2.5 Flash	83/96; 52/96	82/92; 52/92	79/88; 50/88
Mistral Small 3.2 24B	60/96; 35/96	59/92; 34/92	55/88; 31/88
GPT-4.1 mini	72/96; 55/96	70/92; 53/92	66/88; 49/88
Qwen3 30B A3B Instruct 2507	75/96; 28/96	73/92; 28/92	69/88; 26/88

Разрыв между совпадением полей и полнотой набора источников сохраняется во всех трех составах. У Qwen он составляет 47, 45 и 43 задания соответственно; у DeepSeek — одно задание в каждом составе. Это описательная закономерность данного корпуса.

Разбор зарегистрированных ответов

Первый случай выбран как первый по идентификатору зарегистрированный случай внешнего оформления Claude, второй и третий иллюстрируют конкретные ограничения схемы и эталона.

Правильный объект внутри внешнего оформления

В arithmetic-01 источники S2 и S3 сообщают одинаковое поступление 26 и 19 единиц, но разное списание: 16 и 17. При равной силе источников эталон требует conflict, пустое значение и ссылки S2, S3. Claude Haiku 4.5 вернул приведенный ниже объект внутри блока кода, затем добавил пояснение. Здесь воспроизведен только сам блок; наличие последующего пояснения отмечено отдельно.

````
```json
{
  "status": "conflict",
  "value": null,
  "citations": ["S2", "S3"],
  "confidence": 85
}
```
````

Строгий прием отклоняет весь ответ из-за оболочки и последующего текста. Дополнительное правило извлекает единственный объект и засчитывает правильные поля и ссылки. Это случай потери на представлении относительно зафиксированного эталона. Содержательная оговорка: начальный запас в документах явно не назван; эталон исходит из полного учета перечисленных поступлений и списания. Поэтому arithmetic-01 также входит в более консервативный дополнительный расчет с исключением шаблона arithmetic-t4.

Правильная дата при недопустимых ссылках

В deadline-02 GPT-4.1 mini указал эталонную дату, однако представил ссылки числами вместо строковых идентификаторов документов:

```
{
  "status": "answer",
  "value": "2026-04-04",
  "citations": [1, 2],
  "confidence": 95
}
```

Объект читается как JSON, но схема нарушена: ожидаются S1 и S2, а получены числа 1 и 2. Фиксированное извлечение внешнего блока эту ошибку не исправляет. Дата совпадает с эталоном, поэтому называть такой результат выдуманной датой было бы неверно. Для автоматического приема это ошибка контракта ссылки, которую следует регистрировать отдельно.

Пограничная формулировка самого задания

В arithmetic-02 документ S2 сообщает о девяти комплектах по 12 единиц каждый, а вопрос просит общую стоимость комплектов. Эталон задан как 108. DeepSeek V3.2 вернул:

```
{
  "status": "insufficient",
  "value": null,
  "citations": [
    "S1",
    "S2",
    "S3"
  ],
  "confidence": 100
}
```

Замороженная оценка считает этот ответ неправильным. Однако формулировка допускает возражение: произведение числа комплектов и единиц определяет количество, а цена единицы явно отсутствует. Поэтому отказ можно объяснить неоднозначностью вопроса. Такая формулировка встречается в четырех заданиях шаблона arithmetic-t3, включая два разрешимых по эталону. Все опубликованные числа сохраняют исходную оценку; обнаруженное ограничение раскрыто после проверки ответов. Следующий корпус должен развести стоимость и количество, пройти независимую содержательную проверку и новый запуск. Постфактум менять эталон ради улучшения показателя отдельной модели было бы методологически неверно.

Наблюдаемая уверенность моделей

После извлечения JSON расчет охватывает различное число допустимых ответов. Показатель Брайера сопоставляет confidence/100 с правильностью статуса и значения; он не оценивает полноту источников. ECE рассчитан по пяти фиксированным интервалам.

Модель	Включено / исключено	Брайер	ECE, пять интервалов
Claude Haiku 4.5	96 / 0	0.202	0.145
DeepSeek V3.2	95 / 1	0.040	0.020
Gemini 2.5 Flash	91 / 5	0.088	0.088
Mistral Small 3.2 24B	74 / 22	0.348	0.358
GPT-4.1 mini	76 / 20	0.062	0.033
Qwen3 30B A3B Instruct 2507	96 / 0	0.311	0.326

У Gemini все 91 допустимых ответа имели уверенность 100, хотя правильными по полям оказались 83. У Qwen в верхнем интервале находились 77 ответов со средней уверенностью 99,7 % и правильностью 75,3 %. Одновременно в нижнем интервале Qwen девять из десяти ответов были правильными. Это показывает, что часть моделей может трактовать уверенность как уверенность в наличии содержательного значения, а не в правильности всего реше-

ния, включающего отказ. Перед использованием confidence для автоматического маршрута необходимо явно определить его смысл и проверить на отдельной выборке.

Материалы

Запуск выполнен 11 ноября 2026 года через API с регистрацией маршрута обслуживания. Для некоторых моделей журнал содержит нескольких обслуживающих поставщиков, поэтому результат относится к зарегистрированной конфигурации доступа. Случайное назначение конфигураций поставщика в эксперименте не проводилось.

Контрольная сумма SHA-256 корпуса: `4a05f1926cd9870b458bc29c18404a8dc4dc82b150ae50baa3c1b46291c169e2`. Она связывает приведенные идентификаторы с конкретной версией исходных документов.

Порядок цитирования

Центр Поспелова. Ответы языковых моделей по русскоязычным документам. 11 ноября 2026 года. Протокол CP-RU-DOC-2026-01. Электронная публикация и архив воспроизводимости.

Замечания к данным и оцениванию принимаются по адресу expertise@pospelov.center. Для проверки следует указать идентификатор задания, модель, условие запуска и спорное правило оценки.