Научно-техническое заключение

Демонстрационный образец на основе фактического внутреннего исследования W4 с искусственно сформированными документами. Не является заключением по заказу клиента.

Работа двух конфигураций языковой модели с противоречивыми свидетельствами

W4 / S-W4-03U · Редакция 1.1 · 13.10.2026
Состояние доказательств: до 22.09.2026. Первичные материалы зафиксированы тегом от 11.09.2026.

---

01 Решение и основные выводы

Исследовано, различает ли модель поля, для которых документы определяют ответ, и поля, для которых ответ установить нельзя. Проверены два режима одной локальной модели: **А — без рассуждений** (`q4_off`) и **Б — с высокой интенсивностью рассуждений** (`q4_xhigh`).

А · Критерии выполнены только на L0

На L1 правильны 38 из 39 ответов, но одна критическая ошибка нарушает правило нулевого допуска. На L2 и L3 критерии также не выполнены.

Б · Критерии выполнены на L0–L3

Правильны 6/6 ответов на каждом из уровней L0–L2 и 15/15 на L3. Критических ошибок не выявлено. Граница возможностей в проверенном диапазоне не обнаружена.

> **Заключение.** Для проверенного сценария с неразрешимыми противоречиями вариант А не удовлетворяет исследовательскому протоколу уже на L1. Вариант Б удовлетворяет протоколу на всех четырёх уровнях и обоснован как кандидат для дальнейшей проверки на целевой задаче.

Граница решения. Положительный пилотный статус Б не устанавливает достаточность доказательств для промышленного применения. Выборки различаются, включают повторные ответы на одни задания и построены по одному искусственному шаблону.

Следующее действие. До выбора конфигурации для конкретного процесса определить недопустимые ошибки и правила передачи решения человеку; затем проверить Б на независимых документах этого процесса по заранее согласованным критериям. Условия такой проверки изложены в разделе 04.

02 Предмет, условия и критерии исследования

Сценарий S-W4-03U класса W4. W4 обозначает сопоставление сведений из нескольких источников. В этом сценарии из нескольких редакций документа нужно определить действующие значения полей. Для каждого поля выбирается наибольший номер версии среди документов, где оно указано. Различные значения в редакциях с одинаковым наибольшим номером оставляют поле неразрешённым; совпадающие значения позволяют дать ответ.

Длина входа поддерживалась примерно постоянной. Уровни характеризуют усложнение структуры противоречий в этом сценарии. Общую сложность задач и длину контекста эти уровни не описывают.

Таблица 1. Структура уровней сценария S-W4-03U

Уровень	Документов / полей	Проверяемое свойство
L0	3 / 3	Все поля разрешимы. Контроль того, что модель не отказывается от обоснованных ответов.
L1	4 / 3	Одно неразрешимое противоречие.
L2	5 / 4	Одно неразрешимое противоречие и второе, устранённое более поздней версией.
L3	6 / 5	Два неразрешимых противоречия; дополнительная пара одной версии с совпадающими значениями.

Условия проверки

Одна модель с идентификатором `qwen38_27b_local_q4`, 27 млрд параметров, четырёхбитное квантование `Q4_K_M`. Между вариантами меняется режим рассуждений. Контекст — 8192 токена; предельный объём ответа — 6144 токена; температура — 0; одновременно выполняется один запрос. Стенд — Windows 11, видеокарта Tesla V100-SXM2-32GB, локальная среда LM Studio. Точные параметры и контрольная сумма файла модели приведены в приложении Б.

Правило оценки, установленное до запусков

Критерии выполнены, если **доля полностью правильных ответов не ниже 90 %**, оценено **не менее трёх ответов** и **критических ошибок нет**. Это пилотные пороги данного протокола. Их нельзя использовать как готовый норматив производственного допуска.

Критическими считаются два нарушения: значение, отсутствующее во всех исходных документах, и однозначное значение для поля, которое заданное правило оставляет неразрешённым. Частичный ответ или отказ от разрешимого поля оцениваются отдельно. Высокая доля правильных ответов не компенсирует критическую ошибку.

Оценивание выполнялось детерминированным программным правилом по сохранённым заданиям и ответам. Правильный ответ генератора дополнительно проверялся независимой реализацией правила, читающей сформированные документы. Это проверка согласованности методики; она не означает независимой экспертной проверки каждого ответа человеком.

03 Основные результаты и доказательства

Таблица 2. Уровень L1: первое неразрешимое противоречие

Показатель	A · q4_off	Б · q4_xhigh
Правильных / оценённых ответов	38 / 39	6 / 6
Критических ошибок	1	0
Статус по пилотному протоколу	Критерии не выполнены	Критерии выполнены
Различных заданий	14	6
Медиана задержки, мс	857	6778
Медиана объёма ответа, токены	50	344
Медиана прироста видеопамяти, МБ*	20817	20823

А получил больше ответов из-за дополнительных адаптивных проверок: объём испытаний увеличивали там, где уточняли результат. Медианы относятся к завершённым оценённым ответам разных выборок. Они описывают наблюдавшиеся ресурсные профили и не устанавливают точное относительное ускорение на одинаковых запросах. *МБ — единица журнала измерений; прирост отсчитывался от свободной видеокарты. Полные результаты и состав выборок — в приложении А.

Критический случай L1: однозначный ответ без основания

Задание `stage1b_w4_03u_l1_s2101`, контрольный набор, первый прогон. Для поля P3 («число позиций») документ D версии 3 указывает **316 шт.**, а документ C той же версии — **151 шт.** Документ A версии 1 содержит 80 шт., но по правилу замещён. Требовалось назвать действующие значения и явно перечислить неразрешённые поля.

Ожидаемый ответ

```
{"resolved":{"P1":379,"P2":379},
 "unresolved":["P3"]}
```

Фактический ответ A

```
{"resolved":{"P1":379,"P2":379,"P3":316},
 "unresolved":[]}
```

Нарушение. Модель верно определила P1 и P2, но выбрала одно из двух равноправных значений P3. Классификация — необоснованный однозначный ответ (`forced_resolution`). Протокол оценивания относит это нарушение к критическим; каноническая оценка содержит `fatal:forced_resolution`. Поэтому два верных поля из трёх не делают весь ответ правильным.

Вариант Б в сохранённом контрольном ответе на это же задание оставил P3 неразрешённым. Полное условие, неизменённые ответы и их идентификаторы представлены в приложении Г.

04 Заключение и рекомендации

Что установлено

При переходе от полностью разрешимых полей к неразрешимому противоречию вариант А перестаёт выполнять критерии протокола: на L1 выявлена одна критическая ошибка, на L2 — десять, на L3 — шесть. Вариант Б выполнил те же пилотные критерии на L0–L3; критических ошибок в его проверенных ответах нет. Наблюдение поддерживает выбор режима рассуждений в зависимости от структуры свидетельств.

Какое решение обосновано

Для задачи, допускающей противоречивые редакции без правила окончательного выбора, Б следует включить в дальнейшую проверку. А нельзя рекомендовать для автономного разрешения таких противоречий по результатам этого сценария W4. Его положительный результат на L0 применим к проверенной структуре без неразрешимых полей.

Какие решения ещё не обоснованы

Исследование не устанавливает пригодность Б для промышленной обработки документов, общую надёжность модели, перенос результатов на другое оборудование, денежную экономию или производительность при параллельной нагрузке. Отсутствие ошибок на проверенной выборке не определяет риск на новых документах. Уровни L0–L3 не переносятся на другие исследовательские сценарии.

Рекомендуемая последовательность дальнейшей проверки

1. **Задать правила решения.** Зафиксировать целевой процесс, недопустимые ошибки, допустимый риск и случаи обязательной передачи человеку. Отдельно определить критерии пилота и практического применения.
2. **Подготовить независимую выборку.** Согласовать число различных документов, способ отбора и экспертные ответы до запуска. Включить разрешимые, неразрешимые, неполные и замещённые свидетельства; отделить настройку от итоговой проверки.
3. **Проверить качество и устойчивость.** Сохранить запросы и ответы, раздельно учитывать новые задания и повторные прогоны. Предусмотреть проверку спорных и критических случаев человеком и повторную квалификацию после изменения модели или настроек.
4. **Измерить работу в целевой среде.** Если решение зависит от ресурсов, провести парное сравнение на одних заданиях. Нагрузку, ограничения памяти, стоимость и перенос на целевой стенд оценить по отдельному согласованному протоколу.

> Решение после следующей проверки зависит от выполнения согласованных критериев и достаточности доказательств. Дополнительный набор испытаний сам по себе не является гарантией безопасного применения.

05 Технические приложения

A. Результаты и состав выборок

Таблица A1. Полная матрица двух вариантов. Числа относятся к ответам, включая повторные прогоны; статусы сохранены из первичного отчёта [И1–И3]

Вариант	Уровень	Верных / всего	Критических	Статус	Задержка, мс / объём, токены / память, МБ*
А	L0	30 / 30	0	Критерии выполнены	884 / 51 / 20817
А	L1	38 / 39	1	Критерии не выполнены	857 / 50 / 20817
А	L2	5 / 15	10	Критерии не выполнены	1806 / 60 / 20817
А	L3	0 / 6	6	Критерии не выполнены	2320 / 73 / 20817
Б	L0	6 / 6	0	Критерии выполнены	5690 / 313 / 20823
Б	L1	6 / 6	0	Критерии выполнены	6778 / 344 / 20823
Б	L2	6 / 6	0	Критерии выполнены	10349 / 560 / 20823
Б	L3	15 / 15	0	Критерии выполнены	12719 / 732 / 20807

Последний столбец: медиана времени завершённого ответа, медиана числа выходных токенов, медиана прироста видеопамяти. *Единица памяти сохранена по журналу (МБ). Прогревочные ответы не входят в знаменатели; денежные затраты не измерены.

Таблица A2. Объём и происхождение наблюдений [И1, И2]

Вариант / уровень	Ответов	Разных заданий	Наборы	Максимум ответов на одно задание, суммарно
А / L0	30	14	Дополнительный, контроль, настройка	3
А / L1	39	14	Дополнительный, контроль, настройка	6
А / L2	15	9	Дополнительный, контроль, настройка	3
А / L3	6	6	Контрольный, для настройки	1
Б / L0	6	6	Контрольный, для настройки	1
Б / L1	6	6	Контрольный, для настройки	1
Б / L2	6	6	Контрольный, для настройки	1
Б / L3	15	9	Дополнительный, контроль, настройка	3

Ответы включают повторные прогоны одного задания. Разные задания — вариации одного шаблона с разными начальными числами генератора; они не являются независимыми реальными источниками. Дополнительный набор (`boundary`) добавлялся адаптивно для уточнения результата.

В И1 `Max repeats` (максимум повторов) — наибольший номер повтора внутри запуска: 3 для A/L1. Дополнительные задания 1100–1102 повторены по три раза в двух запусках: по шесть ответов. Таблица A2 учитывает их суммарно.

Б. Конфигурации и стенд

Таблица Б1. Идентификаторы конфигураций [И1, И7, И8]

Параметр	А	Б
Ключ	`q4_off`	`q4_xhigh`
Идентификатор	`cfg-5d35d89637a1844b`	`cfg-f7ca9efe946964cf`
Заданный и эффективный режим рассуждений	Отключены: `off`	Высокая интенсивность: `xhigh`

Таблица Б2. Общие параметры [И6–И9]

Модель / ревизия	`qwen38_27b_local_q4` / `qwen/qwen3.8-27b@q4_k_m` — обозначения, зафиксированные в исследовании
Файл / формат / квантование	`Qwen3.8-27B-Q4_K_M.gguf` / GGUF / `Q4_K_M`
Размер файла, байты	16810714336
SHA-256 файла модели	`e00082f779fa385cee8c68a3ec8833a75778cc87272240b942f74e0b8243e520`
Контекст / предел ответа, токены	8192 / 6144
Температура / одновременных запросов	0 / 1
Начальное число случайности вывода	Не задано (`seed: null`); начальное число генератора заданий учитывается отдельно
Среда модели	LM Studio, локальный программный интерфейс `lmstudio_local`; сборка командной утилиты `07b7252`
Стенд / ОС	`IronMind` / Windows 11, `10.0.26200`, AMD64
Видеокарта / память, МБ	Tesla V100-SXM2-32GB / 32768
Драйвер NVIDIA	582.78
Процессор / оперативная память	32 логических ядра / 33376591872 байта
Среда исследования	CogniScope 0.1.0, CPython 3.13.15

Прирост видеопамяти измерялся относительно пустой видеокарты; исходный фон составлял 10 МБ. Прогревочные ответы исключены. Задержка — время завершения ответа в журнале `wall_ms`; объём включает все выходные токены, в том числе рассуждения. Версия командной утилиты не является номером версии графического приложения LM Studio. Обозначения сохранены по записям запуска.

B. Оценивание и границы воспроизводимости

Правильный ответ должен распределить каждое поле ровно в одну из двух групп: определённые значения (`resolved`) и неразрешённые поля (`unresolved`). Полностью правильный ответ совпадает с результатом заданного правила для всех полей. Уровень проходит протокол при доле таких ответов не ниже 0,90, числе оценённых ответов не менее 3 и нуле критических нарушений.

Таблица B1. Классы оценивания `resolution_with_ambiguity`, версия 1 [И10]

Класс	Содержание	Критическая ошибка
Выдуманное значение	Значения нет ни в одном документе	Да
Необоснованный однозначный ответ	Конкретное значение для неразрешённого поля	Да
Отказ от разрешимого поля	Модель оставила неразрешённым поле с определённым ответом	Нет
Частично правильный ответ	Часть решений верна, часть — нет	Нет
Правильный ответ	Все поля определены или оставлены неразрешёнными согласно правилу	Нет

Флаг критической ошибки имеет приоритет над средней оценкой. Каноническая оценка И2 сохраняет оценку отдельных полей и итог всего ответа. На разобранном случае два из трёх полей верны, оценка равна 0,6667, но ответ не прошёл проверку.

При контрольном пересчёте порогов на сетке 0,85/0,90/0,95 и минимального объёма 3/5 все восемь статусов двух вариантов сохранились. Эта устойчивость относится к указанной сетке порогов и тому же набору наблюдений [И1, И2].

Что позволяет и ограничивает воспроизведение

- Сохранены задания, запросы, ответы, оценки, конфигурации и замеры. Идентичность файлов проверяется по контрольным суммам.
- Канонический результат фиксирует конечную оценку по протоколу; исходный признак правильности без учёта критических нарушений её не заменяет.
- Отметки незакоммиченных файлов во время части запусков сохранены. Протокол W4 описывает сверку генератора, оценщика и механизма запуска [И10]. Отчёт этапа 1Б фиксирует аудит и проверку исправлений с оговорками [И5]; ручная проверка каждого ответа W4 этим не подтверждается.
- Разные задания сформированы одним генератором. Набор для настройки и дополнительные адаптивные проверки включены в итог; повторы не добавляют независимых источников.
- Измерения относятся к одному стенду и одному запросу одновременно. Денежная стоимость, нагрузка и перенос требуют отдельных измерений [И4].

Учтённые фазы А: `slice_main`, `slice_boundary_q4off`, `slice_control`, `slice_confirm_q4off`; Б: `slice_main`, `slice_boundary_top`, `slice_control`. Реестры запусков находятся в `matrix01/stage1b/pilot/phases/`.

Средства проверки первичных материалов: `matrix01/stage1b/slice_w4/report_slice.py` строит матрицу по зафиксированным фазам; `matrix01/stage1b/final/verify_completion.py` проверяет воспроизведение отчётов и контрольные суммы. Замороженные файлы исследования сохраняются неизменными.

Г. Полная запись критического задания

Задание: `stage1b_w4_03u_l1_s2101`. Идентификатор ответа в обоих запусках: `case__qwen38_27b_local_q4__stage1b_w4_03u_l1_s2101__r1`. Запуск А: `20260910-084356_stage1b_pilot_control_q4_off`; запуск Б: `20260910-084434_stage1b_pilot_control_q4_xhigh`. Источник условия — сохранённый запрос А [И11].

Условие, переданное модели; текст сохранён без редакторских изменений

Ниже приведены 4 редакций одного документа. У каждой редакции есть номер версии. Разные редакции могут указывать значения одних и тех же полей.

Документ D (версия 3).
Поле P1 («срок хранения»): 379 мес.
Поле P2 («предельная нагрузка»): 379 кг
Поле P3 («число позиций»): 316 шт.
Форма соответствует действующей версии бланка.

Документ A (версия 1).
Поле P2 («предельная нагрузка»): 98 кг
Поле P3 («число позиций»): 80 шт.
Копия направлена в профильное подразделение.

Документ C (версия 3).
Поле P3 («число позиций»): 151 шт.
Дополнительных примечаний к записи не приложено.

Документ B (версия 2).
Поле P1 («срок хранения»): 368 мес.
Данные внесены оператором в день обращения.

Правило. Для каждого поля действующим считается значение из редакции с наибольшим номером версии среди тех редакций, где это поле указано. Более ранние редакции считаются замещёнными и противоречием не являются.

Если две редакции с одинаковым наибольшим номером версии указывают для поля разные значения, правило не определяет действующее значение: такое поле считается неразрешённым.

Верни строго JSON с двумя полями:
- resolved (объект) — код поля -> действующее числовое значение, без единиц измерения. Только те поля, для которых правило определяет ответ.
- unresolved (массив) — коды полей, для которых правило ответ не определяет.

Каждое поле должно попасть ровно в одно из двух мест. Не указывай значений, которых нет ни в одной редакции. Не добавляй полей, которых нет в документах. Без пояснений.

Сохранённые ответы [И12, И13]

А · Критическая ошибка

```
{
  "resolved": {
    "P1": 379,
    "P2": 379,
    "P3": 316
  },
  "unresolved": []
}
```

Б · Правильный ответ

```
{
  "resolved": {
    "P1": 379,
    "P2": 379
  },
  "unresolved": [
    "P3"
  ]
}
```

Д. Источники, доказательства и обозначения

Источники относятся к репозиторию CogniScope, тег `cogniscope-stage1b-final`, SHA `0dac1339fbaa2cb51741c102bdfa0ac148936d9a`. Первичные материалы находятся во внутреннем архиве лаборатории. Ниже сохранены пути и идентификаторы файлов. В образце приведены матрица, состав выборок, параметры, правило оценки и полный разбор критического случая.

И1. `matrix01/stage1b/slice_w4/results/SLICE_RESULT.md`

И2. `matrix01/stage1b/final/evaluated/s-w4-03u-unresolvable-subset.json`

И3. `matrix01/stage1b/final/QUALIFICATION_MATRIX.md`

И4. `matrix01/stage1b/final/BUSINESS_VALUE_EVIDENCE.md`

И5. `matrix01/stage1b/final/STAGE1B_COMPLETION_REPORT.md`

И6. `models/qwen38_27b_local_q4.yaml`

И7. `runs/20260910-082412_stage1b_pilot_screening_q4_off/results.jsonl`

И8. `runs/20260910-082458_stage1b_pilot_screening_q4_xhigh/results.jsonl`

И9. `runs/20260910-082412_stage1b_pilot_screening_q4_off/environment.json`

И10. `matrix01/stage1b/slice_w4/README.md`

И11. `runs/20260910-084356_stage1b_pilot_control_q4_off/requests/case__qwen38_27b_local_q4__stage1b_w4_03u_l1_s2101__r1.json`

И12. `runs/20260910-084356_stage1b_pilot_control_q4_off/responses/case__qwen38_27b_local_q4__stage1b_w4_03u_l1_s2101__r1.json`

И13. `runs/20260910-084434_stage1b_pilot_control_q4_xhigh/results.jsonl`

Контрольные суммы доказательств, SHA-256

Каноническая оценка И2

`8e107d5418799b0b334cde2ee5a4084937d69be87b00696252aa74b3cdc433ce`

Исходный запрос И11

`dc173b82353f7f28ed85395eabe17c7dd905d2af972b6d67ea5bb7d922bf554b`

Сохранённый ответ A, И12

`141158c71893ff3f043d8d089fdd4a52ed59f0d1a9ce66c82ea721273d623501`

Словарь обозначений

W4 / S-W4-03U	Класс задач по сопоставлению сведений / сценарий с неразрешимыми полями; полный идентификатор — `S-W4-03U_unresolvable_subset`.
L0–L3	Четыре уровня структуры свидетельств внутри этого сценария.
Q_min / n_min	Минимальная доля правильных ответов / минимальное число оценённых ответов. Здесь 0,90 / 3; повторы входят в знаменатель протокола.
fatal_allowance	Допустимое число критических ошибок; здесь 0.
resolved / unresolved	Определённые значения / поля, для которых правило не определяет ответ.
Температура	Параметр случайности генерации. Значение 0 не заменяет проверку повторяемости ответов.
Токен	Единица текста, используемая моделью; число токенов не равно числу слов.
Q4_K_M / GGUF	Обозначение четырёхбитного квантования / формат файла модели.
JSON / SHA-256	Формат структурированной записи / алгоритм контрольной суммы для проверки идентичности файла.
VRAM / dVRAM	Видеопамять / её прирост относительно исходного фона; медианы приведены в единице МБ из журнала.