DATANOMIX · OPEN RESEARCH · V1.0

Какая LLM лучше всех пишет *Qlik Set Analysis*?

Бенчмарк 13 больших языковых моделей на 31 верифицированной задаче Qlik Set Analysis из трёх доменов: Sports, HR, Sales. Двухфазная методология, двойной независимый LLM-судья. До 77% решений возвращают верное число — но только 22–34% используют логику, эквивалентную эталонной формуле.

MODELS	TASKS	DOMAINS	BUDGET	UPDATED
13	31	3	$17.35	MAY 2026

TL;DR

Резюме в четырёх пунктах.

ЕСЛИ ВРЕМЯ ПОДЖИМАЕТ, ЧИТАЙТЕ ЭТО.

01 Протестировали **13 LLM-моделей** на 31 задаче Qlik Set Analysis из 3 разных доменов (Sports, HR, Sales). Задачи реальные, с эталонными ответами и автопроверкой.

02 Использовали **двухфазную методологию** + проверку стабильности + двойную проверку правильности (по числовому ответу и по логике выражения).

03 Разница между «по числу» и «по логике» огромная: **65–77%** у топ-моделей по числу — но только **22–34%** по строгой логической эквивалентности. Часть «правильных» ответов — совпадение, которое не обобщится на других данных.

04 Production-вывод: использовать LLM только с обязательной валидацией результата человеком или Qlik runtime-ом. Лучшая модель — **GPT-5** — даёт ~34% строго-правильных. Бюджет **$17.35 из $20**.

Ключевые цифры

MODELS TESTED	TASKS	NUMBER-MATCH	LOGIC-MATCH
13	31	65–77%	22–34%
OPENAI · ANTHROPIC · GOOGLE · META · …	VERIFIED SET ANALYSIS	TOP TIER	TOP TIER · STRICT

ЦЕЛИ ИССЛЕДОВАНИЯ

Четыре цели.

01. **Понять**, какие LLM-модели реально справляются с генерацией Qlik Set Analysis.

02. **Сравнить** модели по точности, стоимости, скорости и стабильности.

03. **Проверить гипотезу**: можно ли промпт-инжинирингом дешёвую модель довести до уровня дорогой.

04. **Сформировать** data-driven рекомендации для возможной интеграции LLM в продукт.

МЕТОДОЛОГИЯ

Двухфазная схема, двойной судья.

Задачи – с обучающей платформы qata.datanomix.pro. Реальные, с эталонными выражениями и автопроверкой результата. Никаких выдуманных исследователем кейсов.

Источник задач

Три набора задач, отобранных по сложности:

- **Sports.Set Analysis Initiate** — 13 простых задач (Олимпиады).
- **HR.Set Analysis Master** — 10 сложных задач (зарплаты сотрудников).
- **Tensini Challenge.Set Analysis** — 8 средних задач (продажи).

Итого **31 уникальная подзадача** из 3 доменов и 3 уровней сложности. Платформа доступа: **OpenRouter** (единый API к 300+ моделям), бюджет $20.

Phase 1 + Phase 2

PHASE 1

13 моделей × 31 задача × 1 промпт

Отбор. Каждая из 13 моделей решает все 31 задачу с одним стандартным промптом. На выходе — leaderboard по двум проверкам и шорт-лист топ-5 моделей.

PHASE 2

5 финалистов × 3 промпта

Топ-5 моделей × 31 задача × 3 уровня промпта (минимальный / стандартный / обогащённый). Цель — измерить эффект промпт-инжиниринга.

Двойной независимый судья

Каждый ответ модели прогоняли через двух LLM-судей. Один смотрел *что получилось*, второй — *как это написано*. Когда расходятся — появляется «логический разрыв».

ПРОВЕРКА №1 · CLAUDE OPUS 4.7

«Совпало ли итоговое число с эталонным KPI?»

Судья запускает выражение модели в Qlik и сверяет полученное число с эталонным KPI из тренинговой платформы. Если число совпало — засчитано, логика выражения не анализируется.

`Топ-модели: 65–77%`

ПРОВЕРКА №2 · CLAUDE SONNET 4.6

«Эквивалентно ли выражение эталонной формуле?»

Судья сравнивает Set Analysis-выражение с эталонным с qata.datanomix.pro. Засчитано только если выражения семантически эквивалентны. Совпало число «случайно» через другую логику — не засчитано.

`Топ-модели: 22–34%`

КАНДИДАТЫ

13 моделей · 4 категории.

Не брали устаревшие версии (Llama 2, GPT-3.5), variant fine-tunes (для roleplay/медицины), мелкие модели (≤8B параметров).

КАТЕГОРИЯ	МОДЕЛИ	ОБОСНОВАНИЕ
Топ-премиум	Claude Opus 4.7 · GPT-5 · Gemini 2.5 Pro	Флагманы. Проверить оправданность цены.
Средние	Sonnet 4.6 · GPT-5 mini · Gemini 2.5 Flash · Mistral Large · Grok 3	Sweet spot для production.
Бюджетные	Haiku 4.5 · Llama 3.3 70B · Qwen 2.5 72B	Экономия при сохранении качества.
Спец. для кода	DeepSeek Coder V3 · Qwen 2.5 Coder 32B	Может ли специализация на коде дать преимущество.

PHASE 1 · LEADERBOARD

13 моделей, ранжированы по совпадению числа.

Один стандартный промпт × 31 задача. Колонка **Coincidental** – сколько раз модель «угадала» число через выражение, отличающееся от эталона.

#	MODEL	PROVIDER	NUMBER OK	LOGIC OK	COINC.	TIER
01	**Gemini 2.5 Pro**	GOOGLE	**24/31 (77%)**	**13/31 (42%)**	6	TOP
02	**GPT-5**	OPENAI	**24/31 (77%)**	**9/31 (29%)**	9	TOP
03	**Claude Opus 4.7**	ANTHROPIC	**21/31 (68%)**	**9/31 (29%)**	4	TOP
04	**Claude Sonnet 4.6**	ANTHROPIC	19/31 (61%)	9/31 (29%)	5	MID
05	**Grok 3**	XAI	17/31 (55%)	8/31 (26%)	6	MID
06	**Claude Haiku 4.5**	ANTHROPIC	14/31 (45%)	6/31 (19%)	6	MID
07	**DeepSeek V3** LOCAL	DEEPSEEK	13/31 (42%)	6/31 (19%)	3	MID
08	**Mistral Large**	MISTRAL	11/31 (35%)	7/31 (23%)	3	MID
09	**Gemini 2.5 Flash**	GOOGLE	8/31 (26%)	2/31 (6%)	5	LOW
10	**GPT-5 mini**	OPENAI	6/31 (19%)	4/31 (13%)	2	LOW
11	**Qwen 2.5 72B** LOCAL	ALIBABA	6/31 (19%)	3/31 (10%)	5	LOW
12	**Llama 3.3 70B** LOCAL	META	3/31 (10%)	2/31 (6%)	2	LOW
13	**Qwen 2.5 Coder 32B** LOCAL	ALIBABA	4/31 (13%)	1/31 (3%)	2	LOW

* DeepSeek Coder V3 исключён — API broken (0/31).

PHASE 2 · 5 FINALISTS × 3 PROMPTS

Кто держится при варьировании промпта.

Топ-5 моделей × 31 задача × 3 уровня промпта = 93 ответа на модель. Ранжировано по совпадению логики.

МОДЕЛЬ	LOGIC OK (V2)	NUMBER OK (V1)	КОММЕНТАРИЙ
GPT-5	**32/93 (34%)**	51/93 (55%)	ЕДИНСТВЕННЫЙ ЯВНЫЙ ЛИДЕР

МОДЕЛЬ	LOGIC OK (V2)	NUMBER OK (V1)	КОММЕНТАРИЙ
Gemini 2.5 Pro	30/93 (32%)	43/93 (46%)	CLOSE 2ND
Claude Opus 4.7	24/93 (26%)	45/93 (48%)	TOP TIER
Claude Sonnet 4.6	20/93 (22%)	43/93 (46%)	SWEET SPOT
DeepSeek V3	14/93 (15%)	27/93 (29%)	BUDGET

FINDINGS

Шесть технических открытий.

⚠ 4.1 REASONING TRAP

Reasoning-модели нужно настраивать иначе.

При первом прогоне **GPT-5 = 0/31**, **Gemini 2.5 Pro = 2/31**. Эти reasoning-модели тратят токены на скрытое размышление (*thinking*), которое не возвращается пользователю, но расходует тот же лимит токенов.

При `max_tokens=500` весь бюджет уходит на reasoning, и модели возвращали либо пустой ответ (GPT-5), либо обрезанное выражение (Gemini Pro). **Решение**: `max_tokens=4000 + reasoning_effort=low`. После фикса: **GPT-5 → 24/31 (77%)**, **Gemini 2.5 Pro → 24/31 (77%)**.

★ 4.2 COINCIDENTAL CORRECTNESS — ГЛАВНОЕ ОТКРЫТИЕ

Верное число из выражения, не совпадающего с эталоном — 114 случаев.

Из 868 ответов в Phase 1 + Phase 2 нашли **114 случаев**, когда модель вернула верное число, но через выражение с другой логикой. Два типичных паттерна:

Паттерн А · ID вместо Name (Sports task #2):

ЭТАЛОН

```
count(distinct {<Sex={"M"}>} Name)
/ count(distinct Name)
```

LLM (СОВПАЛО СЛУЧАЙНО)

```
Count({<Sex={'M'}>} DISTINCT ID)
/ Count(DISTINCT ID)
```

Совпало потому что в датасете `ID` уникальный per-athlete. На данных где у одного атлета несколько ID — даст другой результат.

Паттерн Б · Games вместо Year+Season (Sports task #1):

ЭТАЛОН

```
{<Year = {'1996'},
   Season = {'Summer'}>}
```

LLM (СОВПАЛО СЛУЧАЙНО)

```
{<Games = {'1996 Summer'}>}
```

Совпало потому что `Games` — конкатенация Year+Season в этом датасете. Не обобщается.

◆ 4.3 НЮАНС

Не все 114 случаев — строго неправильные.

Часть «coincidental» случаев — легитимные альтернативные решения, которые на этих данных дают тот же результат и могут считаться допустимыми в production. Если в схеме `ID` гарантированно уникален per-athlete, `Count(distinct ID)` = `Count(distinct Name)` всегда.

Реалистичная оценка точности — **между «по числу» и «по логике»** интерпретациями.

⚠ 4.4 PROMPT EFFECT · COUNTER-INTUITIVE

Обогащённый промпт ухудшает результаты у средних моделей.

В Phase 2 тестировали 3 уровня промпта: минимальный (только вопрос), стандартный (схема + роль), обогащённый (плюс примеры + best practices + chain-of-thought).

Обогащённый промпт **ухудшил 3 из 5 моделей**: Sonnet, Gemini Pro, DeepSeek V3. Только премиум reasoning-модели (Opus, GPT-5) выиграли от обогащения.

Средние модели «слепо копируют» структуру из примеров few-shot, теряют гибкость на нестандартных задачах.

✗ 4.5 ГИПОТЕЗА НЕ ПОДТВЕРДИЛАСЬ

Умный промпт не превращает дешёвую модель в дорогую.

DeepSeek V3 с обогащённым промптом показал **более низкий** результат, чем со стандартным: V1 45% → 36%, V2 15%.

Гипотеза «дешёвая модель + умный промпт = дорогая» не подтвердилась. Промпт-инжиниринг не сокращает разрыв между бюджетными и премиум моделями.

∿ 4.6 STABILITY NOISE ±5-15 П.П.

Повторный прогон даёт другие числа.

На одинаковых задачах с temperature=0:

GPT-5	23 → 24	+1
Claude Opus 4.7	19 → 23	+4
Gemini 2.5 Pro	19 → 22	+3
Claude Sonnet 4.6	20 → 20	±0 · единственная стабильная
DeepSeek V3	14 → 12	-2

Источники шума: модели не строго детерминированы при temperature=0, плюс LLM-судья тоже даёт разные вердикты. **Утверждения «X лучше Y на 3-5 п.п.» по нашим данным не доказываются** — это в пределах шума.

COST BREAKDOWN

$17.35 на весь бенчмарк.

~4 300 запросов, ~2.7M токенов. 70% бюджета съел LLM-as-judge (Claude Opus в Phase 1) — при повторе с Sonnet стоимость в **14 раз ниже** за то же количество ответов.

МОДЕЛЬ · РОЛЬ	SPEND	REQUESTS	TOKENS
Claude Opus 4.7 · судья V1	$12.30	1,980	1.81M
Gemini 2.5 Pro · кандидат	$1.91	253	247K
GPT-5 · кандидат	$1.46	253	199K
Sonnet 4.6 · кандидат + судья V2	$0.85	870	~150K
Остальные 9 моделей	$0.83	950	320K
Итого	$17.35	~4,300	~2.7M

Подтверждена гипотеза «использовать Sonnet/Haiku в роли судьи» — экономия 5-14× без потери качества оценки.

PRODUCTION GUIDANCE

Если LLM пойдёт в продукт.

Три сценария интеграции с реалистичной точностью (с обязательным человеческим ревью) и стоимостью на 1 000 запросов.

СЦЕНАРИЙ	МОДЕЛЬ	ПРОМПТ	ТОЧНОСТЬ*	$/1000
Базовый ассистент	Claude Sonnet 4.6	стандартный	~30–50%	~$2
Премиум · критич. задачи	GPT-5	стандартный	~35–55%	~$20
Прототипирование	DeepSeek V3	стандартный	~15–30%	~$0.30

* С обязательным человеческим ревью.

PRODUCTION REQUIREMENTS

Четыре правила, без которых не идти в прод.

01. **Никогда без ревью.** Никогда не использовать без человеческого ревью или Qlik runtime-валидации. Лучшая модель даёт ~34% строго-правильных ответов — каждый второй ответ требует проверки.

02. **Настроить reasoning-модели.** GPT-5, Gemini 2.5 Pro требуют `max_tokens=4000` + `reasoning_effort=low`. Иначе систематически заниженные результаты.

03. **Не перегружать few-shot.** Для большинства моделей обогащённый промпт *снижает* точность. Простой промпт + строгая валидация работают лучше.

04. **Sonnet/Haiku в роли судьи.** Не Opus. Экономия 5–14× без потери качества оценки — проверено на 868 ответах.

ON-PREM DEPLOYMENT

Какую open-source модель развернуть локально?

Отдельный вопрос: если LLM в облаке нельзя по политике безопасности – что брать on-prem.

★ LOCAL DEPLOYMENT RECOMMENDATION

Из локальных моделей, которые мы протестировали, лучший — **DeepSeek V3** с **~19%** точности по логике (когда сгенерированное выражение совпадает с эталоном). **Qwen 2.5 72B** заметно хуже — около **10%**. **Qwen 2.5 Coder 32B** вообще слабо — **3%**: для длинных цепочек CALCULATE/SUMX в Set Analysis 32B параметров не хватает. **GLM** мы не тестировали.

Один важный нюанс: даже у лидера правильная логика выражения — **в 1 из 5 случаев**. То есть в продакшене любую open-source модель надо обязательно использовать с валидацией. Без неё пока сыровато.

ЗАКЛЮЧЕНИЕ

Что мы узнали.

Исследование подтверждает: **LLM могут генерировать корректный Qlik Set Analysis** — но с серьёзной оговоркой по строгости оценки. При проверке только по числу — 65–77% точности у топ-моделей. При проверке по эквивалентности логики эталону — 22–34%. Реалистичная оценка для production — между ними, ~30–50%.

> *Главная рекомендация – использовать только в режиме «ассистент для человека», не в режиме автоматической генерации без валидации. Главный технический инсайт – про настройку reasoning-моделей – критически важен для любой команды, которая будет интегрировать GPT-5 / Gemini Pro / o1 / o3 в production.*

Главный методологический инсайт — про двойную проверку (число + логика) — должен стать **стандартом для любых будущих LLM-бенчмарков в команде**.

Краткое резюме по моделям

КРИТЕРИЙ	МОДЕЛЬ	ИНСАЙТ
Лучшая для строгой генерации (V2)	**GPT-5**	Лидер по строгой оценке — 34%.
Базовый ассистент	**Claude Sonnet 4.6**	Sweet spot, ~30–50% (с ревью).
Стоимость Sonnet 4.6 / 1 000 запросов	**~$2**	Экономия до 14× по сравнению с Opus.
Причина выбора Sonnet	**Баланс точности и стоимости**	Приемлемая точность при низкой стоимости.

Reproduce on GitHub: github.com/adinomix/qlik-llm-benchmark