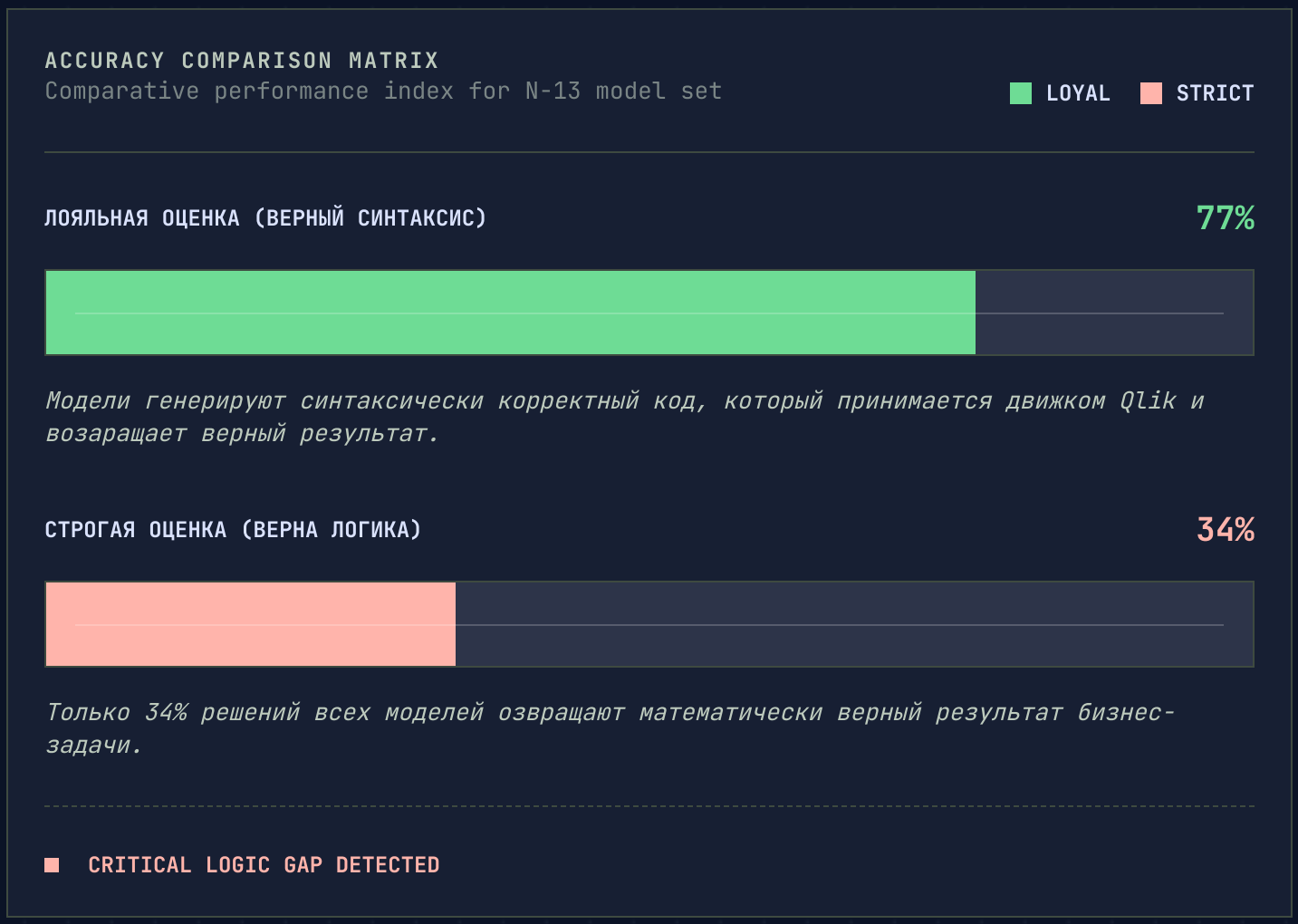
LLM для Анализа Множеств:

Готовность LLM к написанию production-ready Qlik Set Analysis

Бенчмарк 13 моделей на 31 задаче Qlik Set Analysis. Выявлен критический разрыв между лояльной (77%) и строгой (34%) точностью.

ИНСАЙТ

В **77% случаев** LLM генерируют выражения, возвращающие верный числовой результат (лояльная оценка), но лишь 34% используют строгую, эталонную логику. Высокая точность достигается за счёт логически альтернативных, но случайно корректных выражений, которые не гарантируют надёжности.



БЮДЖЕТ

$17.35 / $20.00

ДОМЕН

Advanced Set Analysis в сферах: Sales, HR, Sports

ЗАДАЧ ДЛЯ ПРОВЕРКИ

31

■ МЕТОДОЛОГИЯ

Методология исследования

Мы использовали внутреннюю обучающую платформу qata.datanomix.pro - задачи по Анализу Множеств в Qlik реальные, и содержат эталонные ответы и встроенную автопроверку.

Затем внедрили двухфакторный аудит, объединяющий вычислительную мощность Qlik Engine и экспертизу сертифицированных архитекторов на платформе OpenRouter, с единым API доступом к 300+ моделям. Наш бюджет - $20.

LOYAL JUDGE | CLAUDE OPUS

PHASE_01

ОБЪЕКТ ПРОВЕРКИ:

```
Result Consistency Check
```

Автоматическая сверка итогового значения. Если результат вычисления совпадает с выходным значением в Qlik Engine, ответ считается формально верным.

```
Моделей:                                    13
```

STRICT JUDGE | CLAUDE SONNET

PHASE_02

ОБЪЕКТ ПРОВЕРКИ:

```
Logic & Formula Integrity
```

Глубокий аудит логики выражения. Даже если число совпало случайно, эксперт проверяет использование модификаторов, операторов и соответствие эталонным "верным" выражением.

```
Моделей:                                                    5
Промпт:                          3 варианта (минимальный, стандартный, обогащенный)
```

Выбор моделей — 13 кандидатов

КАТЕГОРИЯ	МОДЕЛИ	ОБОСНОВАНИЕ
Топ-премиум	`Claude Opus 4.7, GPT-5, Gemini 2.5 Pro`	Флагманы, проверить оправданность цены
Средние	`Claude Sonnet 4.6, GPT-5 mini, Gemini 2.5 Flash, Mistral Large, Grok 3`	Sweet spot для production
Бюджетные	`Claude 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 спец. на коде дать преимущество

Технические открытия

1 Ловушка reasoning-моделей

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

```
// Default Settings Failure
max_tokens=500
reasoning_effort=low
```

При дефолтном значении модели тратили весь бюджет на размышления и возвращали либо пустой ответ, либо обрезанное выражение.

```
OPTIMIZED FIX
max_tokens: 4000
reasoning_effort: low
```

77%

Post-fix Accuracy

После установки кастомных значений, количество правильных ответов у топ-2 моделей возросло:

```
GPT-5 DELTA
```

GEMINI DELTA
2/31 → 24/31

2 Модели генерируют альтернативные выражения

Значительная часть правильных ответов получена через выражения, **отличающиеся от эталона**.

114 случаев "правильный ответ из выражения с другой логикой" из 868 ответов

При этом это не считается неправильной логикой. Часть из этих 114 случаев — легитимные альтернативные решения, которые на этих данных дают тот же результат и могут считаться допустимыми в production.

ДЕТАЛИ О ПАТТЕРНАХ В ВЫРАЖЕНИЯХ

PATTERN A 'ID' вместе 'Name'

Эталонное выражение:

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

Ответ модели (успешный):

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

На данных где у одного атлета несколько ID (один на каждое событие) — даст другой результат. В текущем тесте совпало случайно.

* Для контекста, посмотрите задачку #2 в таск-сете Sports.Set Analysis Initiate на qata.datanomix.pro

PATTERN B 'Games' вместо 'Year'+'Season'

Эталонное выражение:

Ответ модели (успешный):

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

Модели используют 'Games' как конкатенацию Year+Season. Логика не обобщается на другие датасеты без этого технического поля.

* Для контекста, посмотрите задачку #1 в таск-сете Sports.Set Analysis Initiate на qata.datanomix.pro

3 Эффект промптов — counter-intuitive

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

Обогащённый промпт ухудшил результаты у 3 из 5 моделей (Sonnet, Gemini Pro, DeepSeek V3).

Только премиум reasoning-модели **(Opus, GPT-5)** выиграли от обогащения. *Итог: средние модели «слепо копируют» структуру из примеров few-shot, теряют гибкость на нестандартных задачах.*

```
SONNET 3.5
Negative Delta
```

```
GEMINI PRO
Negative Delta
```

```
DEEPSEEK V3
Negative Delta
```

4 Гипотеза «дешёвая модель + умный промпт = дорогая» НЕ подтвердилась

Промпт-инжиниринг **не сокращает** разрыв между бюджетными и премиум моделями. Качество базовой архитектуры остается доминирующим фактором в сложных BI-задачах.

DEEPSEEK V3 REGRESSION

45% → 36%

Enriched Prompt Impact

Финальные результаты

Phase 1 — рейтинг 13 моделей

По двум версиям судьи:

МОДЕЛЬ	V1 (ЛОЯЛЬНЫЙ)	V2 (СТРОГИЙ)	COINCIDENTAL CASES
Gemini 2.5 Pro	24/31 (77%)	13/31 (42%)	6
GPT-5	24/31 (77%)	9/31 (29%)	9
Claude Opus 4.7	21/31 (68%)	9/31 (29%)	4
Claude Sonnet 4.6	19/31 (61%)	9/31 (29%)	5
Grok 3	17/31 (55%)	8/31 (26%)	6
Claude Haiku 4.5	14/31 (45%)	6/31 (19%)	6

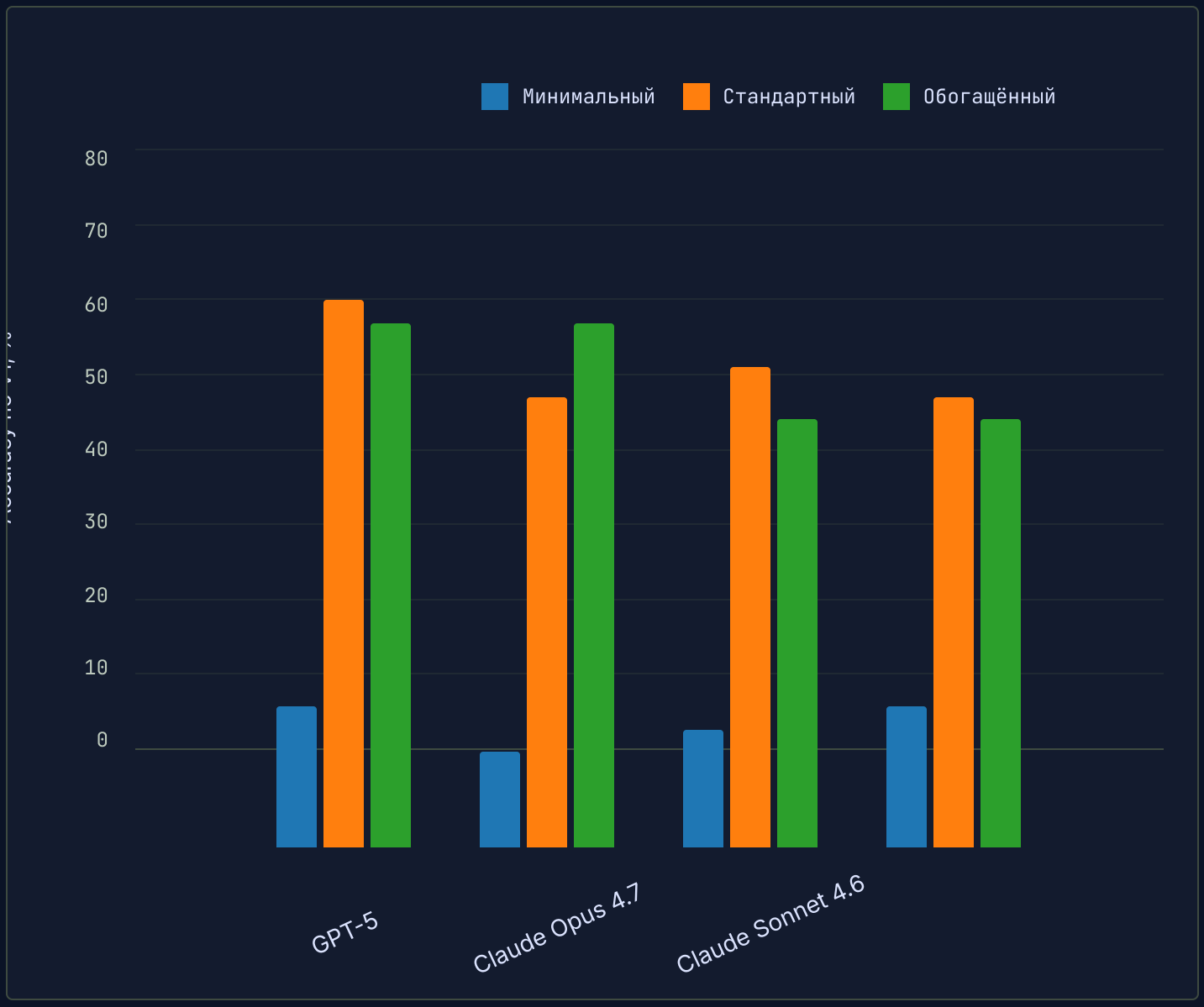
DeepSeek V3	13/31 (42%)	6/31 (19%)	3
Mistral Large	11/31 (35%)	7/31 (23%)	3
Gemini 2.5 Flash	8/31 (26%)	2/31 (6%)	5
GPT-5 mini	6/31 (19%)	4/31 (13%)	2
Qwen 2.5 72B	6/31 (19%)	3/31 (10%)	5
Llama 3.3 70B	3/31 (10%)	2/31 (6%)	2
Qwen 2.5 Coder 32B	4/31 (13%)	1/31 (3%)	2
DeepSeek Coder V3	0/31 (0%)	—	API broken

Phase 2 — топ-5 финалистов с тремя промптами

Точность по V2 (строгий судья) — правильных из 93 (31 задача × 3 промпта):

МОДЕЛЬ	ИТОГО V2	ИТОГО V1
GPT-5	32/93 (34%)	51/93 (55%)
Gemini 2.5 Pro	30/93 (32%)	43/93 (46%)
Claude Opus 4.7	24/93 (26%)	45/93 (48%)
Claude Sonnet 4.6	20/93 (22%)	43/93 (46%)
DeepSeek V3	14/93 (15%)	27/93 (29%)

Phase 2: эффект промптов на accuracy



Заключение

Исследование подтверждает, что LLM могут генерировать корректный Qlik Set Analysis, но с серьезной оговоркой по строгости оценки. Результаты демонстрируют значительный разрыв между формальным сходством и логической эквивалентностью.

Точность у топ-моделей при сравнении только итогового числового результата.

СТРОГАЯ ОЦЕНКА

22-34%

Точность при проверке эквивалентности логики эталонному выражению.

РЕАЛИСТИЧНЫЙ PRODUCTION

~30-50%

Ожидаемый диапазон точности в реальных рабочих сценариях.

Главная рекомендация

Использовать только в режиме **«ассистент для человека»**. Не рекомендуется для режима автоматической генерации без валидации. Предлагайте пользователю выражение для проверки и уточнения, но не для мгновенного применения.

MODEL_NODE / 01 ■ PREMIUM / LOGIC LEADER

Лучшая для строгой генерации: GPT-5

STRICT ACCURACY

34%

Лидер по строгой оценке (эквивалентность логики эталону). Рекомендуется для критических задач, где важна математическая точность.

Оптимальный ассистент: Claude Sonnet 4.6

REALISTIC ACCURACY

~30-50%

COST / 1K REQ

~$2

Баланс точности и стоимости. Рекомендован как базовый ассистент с экономией до 14 раз по сравнению с Opus.