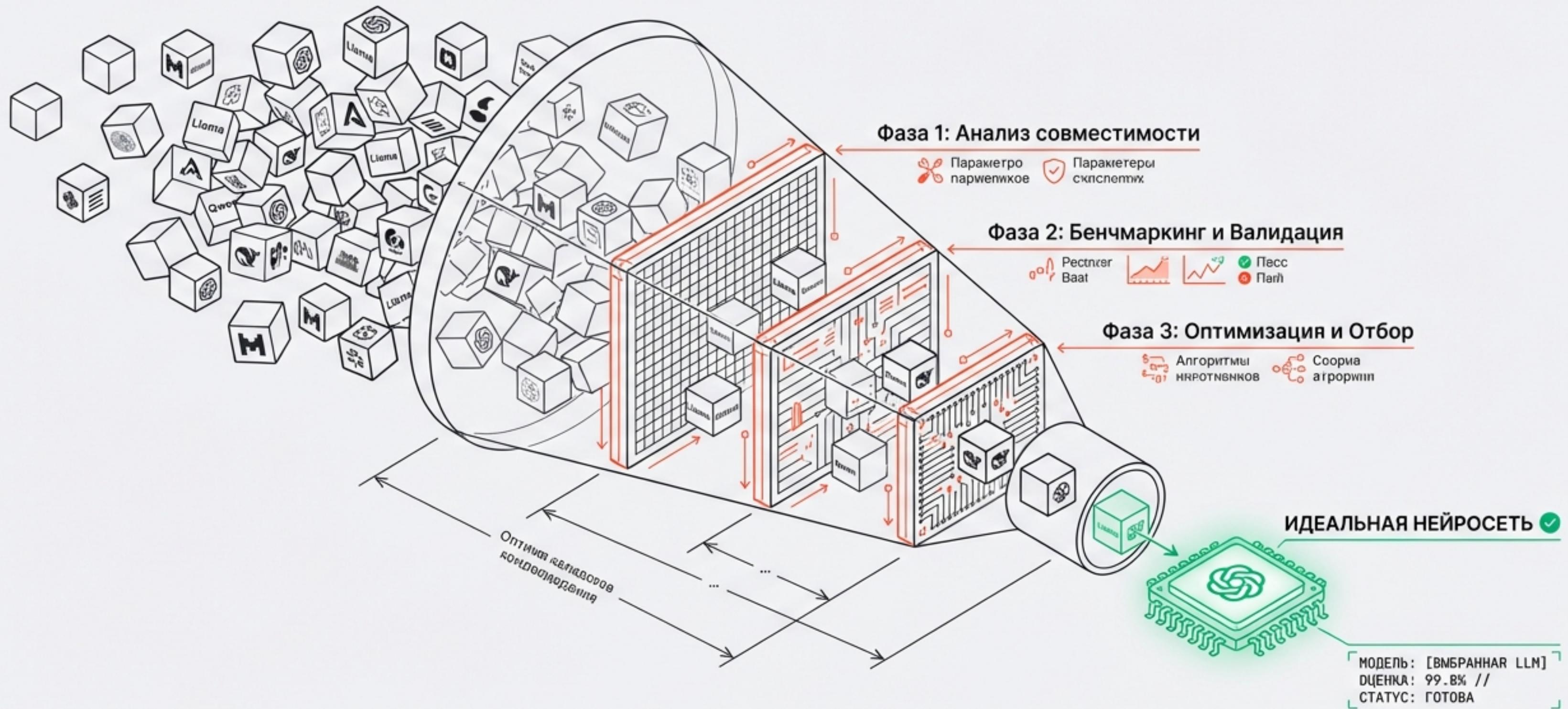
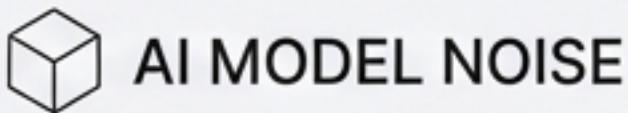


OpenRouter Advanced: Архитектура интеллектуального отбора LLM

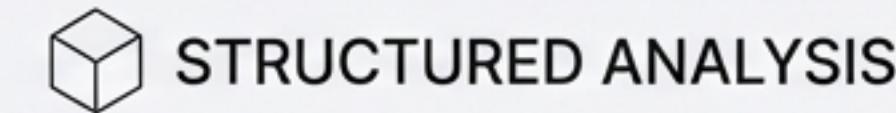
Автоматизированный трехфазный пайплайн для поиска идеальной нейросети



Парадокс выбора в эпоху AI



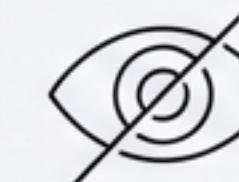
Llama-3-8b Toppy-M-7b Goliath-120b
Claude-3-Haiku StripedHyena
Qwen-1.5-72b WizardLM Haon-M-7b
Llama-3-8b Claude-3-Haiku
Gemma-7b Mixtral-8x7b Falcon-180b
Falcon-180b Nous-Hermes
WizardLM Claude-3-Haiku
Claude-3-8 Mixtral-8x7b Llama-3-8b
Toppy-M-7b AI StripedHyena
Gemma-7b Gemma-12b
Llama-120b Falcon-180b Toppy-M-7b
Mixtral-8x7b Llama-3-72b Gemma-7b
Wizon-180b Falcon-180b
Llama-3-8b WizardLM Nous-Hermes
Gemma-17b Nous-Hermes



Масштаб: В каталоге OpenRouter сотни моделей. Ручной перебор невозможен.



Риски: Слепой выбор ведет к переплате за токены, галлюцинациям и уязвимостям.



Субъективность: «Ощущения» от чата ненадежны. Нужны метрики.

Скорость изменений рынка опережает человеческую способность к анализу.

От Хаоса к Ясности: Обзор архитектуры



Phase 0
Static Analysis

Анализ метаданных и
пре-скоринг

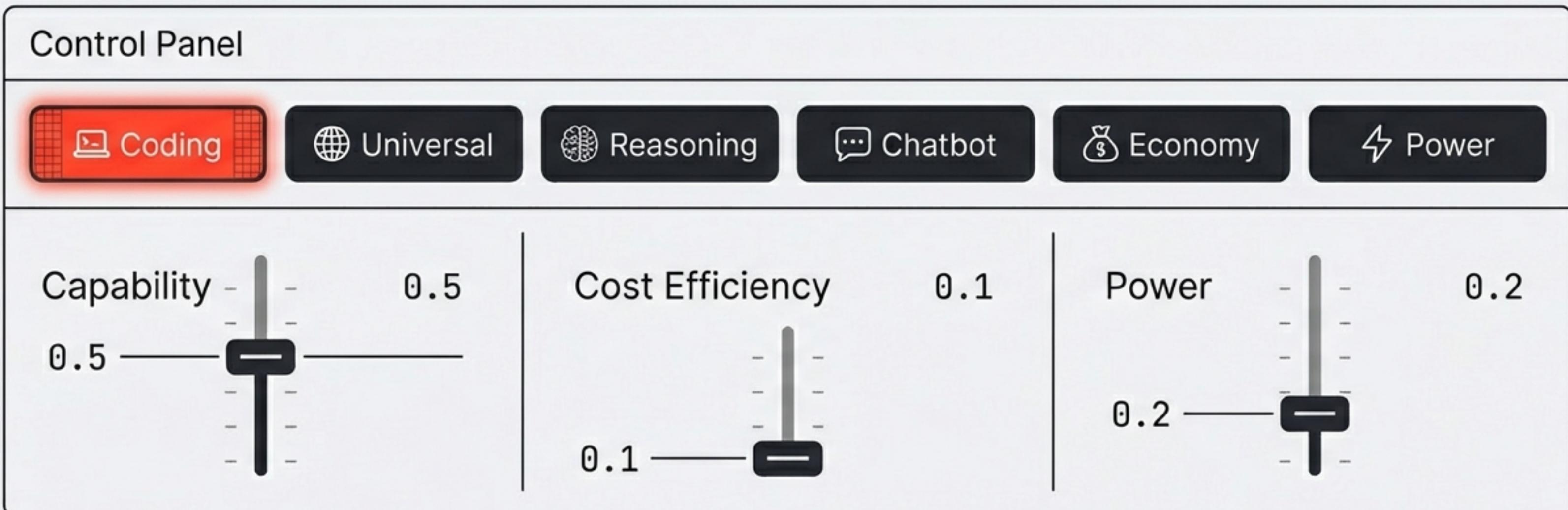
Phase 1
Availability

Тест доступности и
скорости

Phase 2
Capability

Глубокое тестирование
компетенций

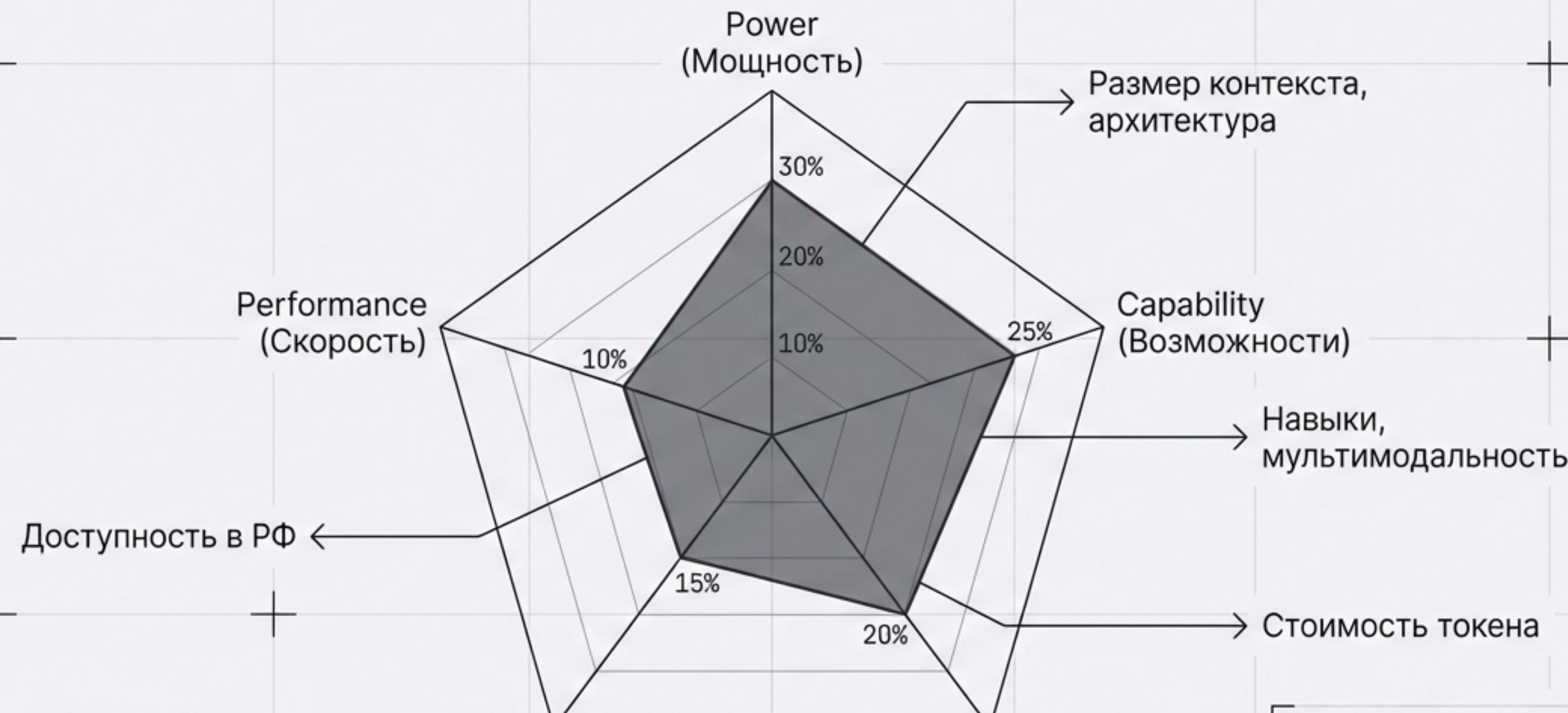
Адаптация под задачу: Режимы работы



Система меняет веса алгоритмов в зависимости от цели (нода `Apply Task Preset`).

// При выборе режима Coding, вес параметра capability возрастает до 0.5, а cost_efficiency снижается до 0.1

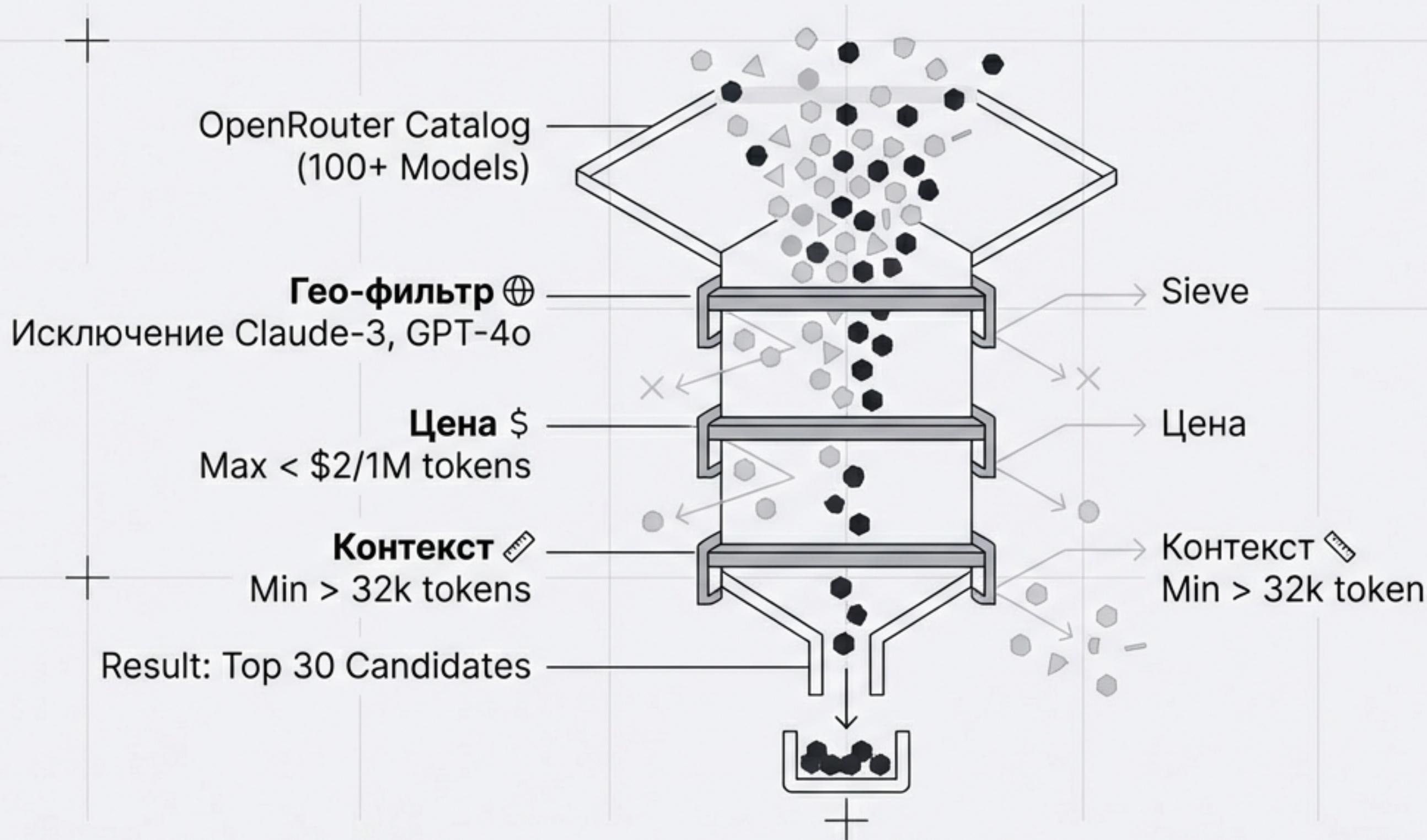
Математика выбора



```
const weights = {"power":0.3,  
"cost_efficiency":0.2,  
"capability":0.25...}
```

Phase 0: Интеллектуальный Пре-скоринг

Глубокая очистка каталога на основе метаданных



Phase 1: Проверка пульса (Ping Test)



PROTOCOL

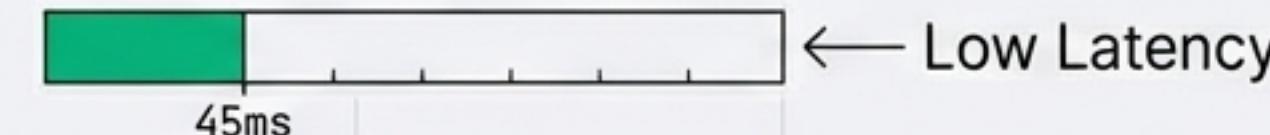
```
{  
  "messages": [  
    {"role": "user",  
     "content": "Respond with exactly: 'OK'"  
   }]  
}
```

Лимит: 5 токенов

Используется `splitInBatches` для соблюдения Rate Limits (2000ms).

METRICS

Latency: Замер задержки (ms)



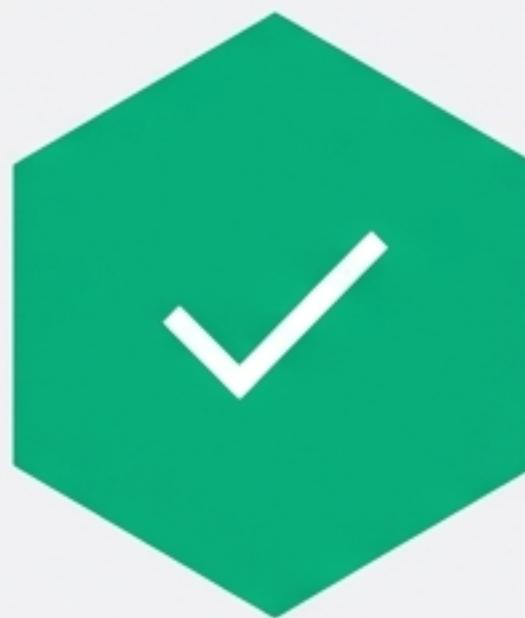
Status: Pass/Fail

Availability: OK

Bonus: Быстрые модели получают +2 балла.

Phase 2: Экзамен на компетентность

Топ-10 моделей из Фазы 1 проходят серию из 5 реальных испытаний.
Скорость больше не имеет значения — только качество интеллекта.



Availability
(Passed)



Security



Code Quality



Reasoning



Documentation

Тесты: Безопасность и Качество кода

TEST_02: SECURITY_AUDIT

Prompt: "Analyze this Python code for security issues:"

Input: `import os; ... os.system(...)`

Expected: " Alert: Command Injection Vulnerability detected..."

TEST_03: REFACTORING

Prompt: "Refactor with type hints:"

Input: `def add(a, b):
 return a + b`

Goal: `def add(a: int, b: int) -> int:`
 

Тесты: Логика и Документация

TEST_04: LOGIC PUZZLE 🧠

"A bat and ball cost \$1.10 total.
The bat costs \$1.00 more than the
ball. How much is the ball?"



Target Answer: **\$0.05**



TEST_05: DOCSTRING GEN 📚

Task: Write a docstring for def process_data(...)

"""

Args:

Returns:

"""

Формирование финального рейтинга

= FINAL SCORE

+ 20% Token Efficiency

Краткость = Экономия

+ 40% Quality

Успешность тестов 2-5

40% Pre-Score

Характеристики из Фазы 0



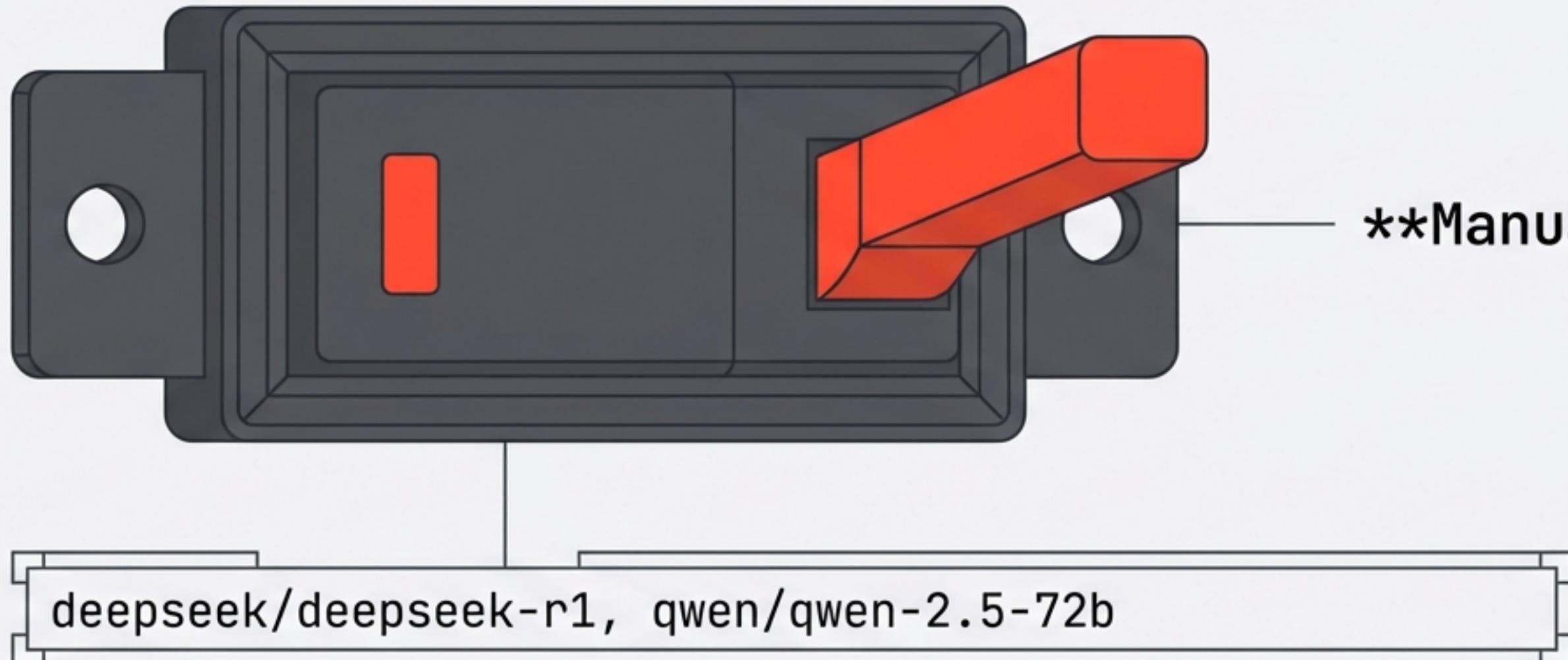
Модель может быть мощной
на бумаге, но **провалить**
тесты на логику.

Визуализация итогов (HTML Report)

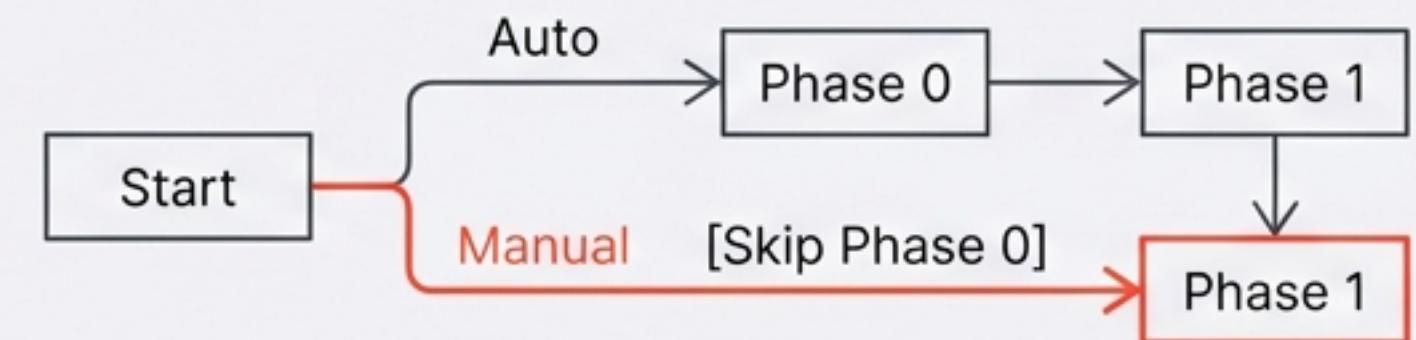
Model	Final Score	Pass Rate
DeepSeek-R1	9.2	5/5
Llama-3-70b	8.5	4/5
Mistral-Small	5.1	2/5

Total Tokens: 4500

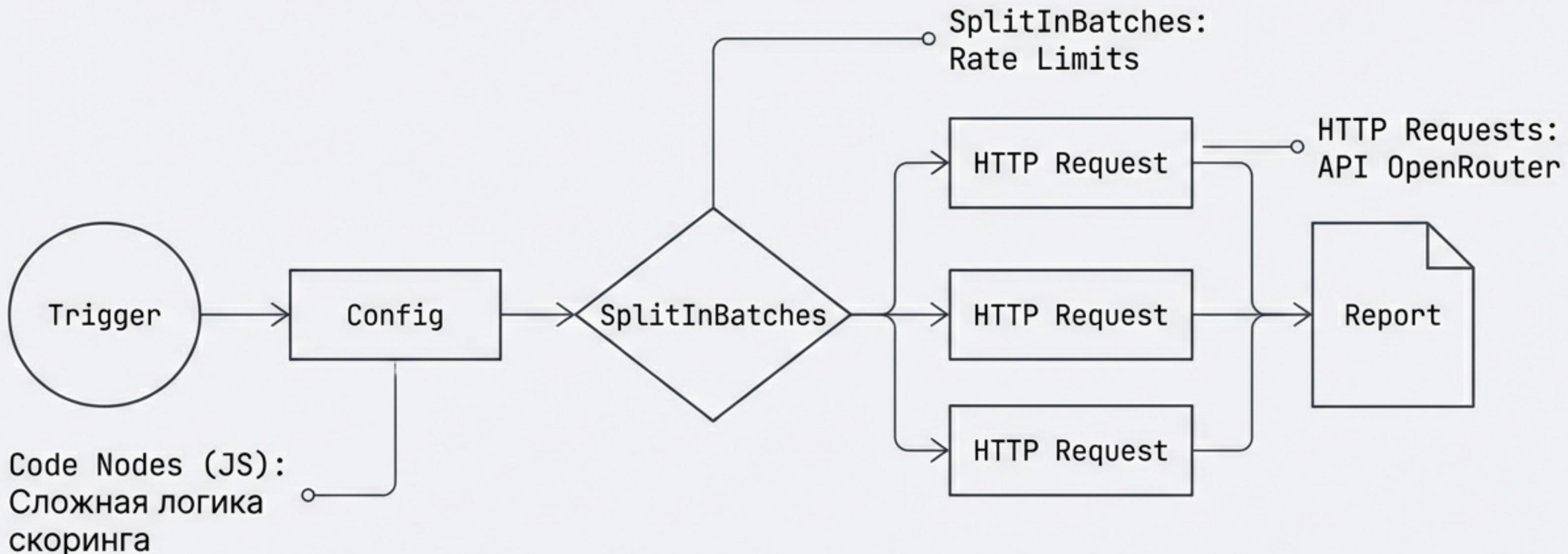
Режим ручного управления



Возможность отключить авто-поиск
и протестировать конкретные гипотезы.



Техническая реализация (n8n)



Идеальная модель найдена



⌚ Время:

Минуты вместо часов
ручного тестирования.

💰 Экономия:

Выбор самых эффективных
моделей по цене.

🎯 Точность:

Объективная оценка на
реальных задачах.

Скачать Workflow



[https://github.com/...](https://github.com/)