

Token FinOps: как инженеры перестают быть расходной статьей

Сколько стоит законченная задача?

Евгений Гребенщиков · Сбер 2026



26,14 ₽

тест-кейсы
на одну кнопку

≈ 1,9 млн ₽

в год
на потоке задач

0

Полностью принятых
Комплектов кейсов

Евгений Гребенщиков

Сбер КИБ трайб DCL

Лидирую тестирование ИИ-агентов

Эксперт профессионального сообщества
тестировщиков Сбера

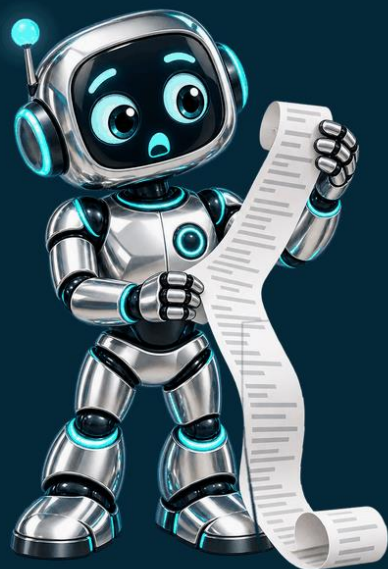
Создаю агентов по разным направлениям



@COOLLight

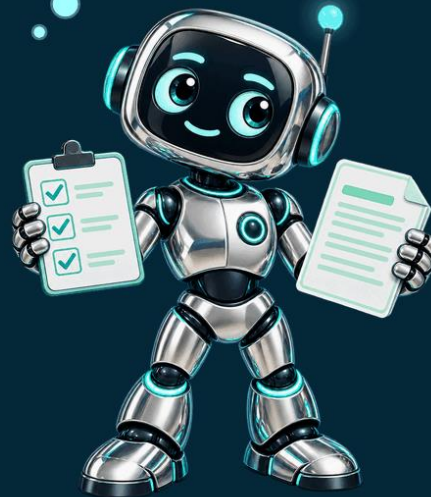


Классно.
А сколько это стоит?

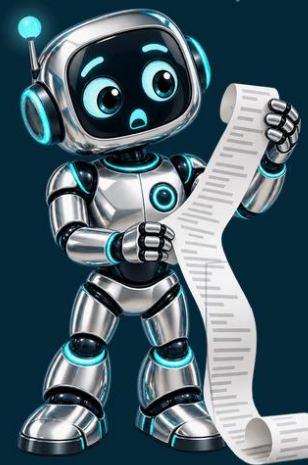


Миллион токенов
стоит 650 ₽...

...а один готовый
результат?



Тогда вы —
расходная статья.



~~ИИ-тестирование - 1,9 млн ₽~~

сократить?

Маршрут

Расчёт

1
Не платить
дважды

2
Не нести
лишнее

3
Не
переплачивать
за модель

Эксперимент
на понедельник

Работает

Требует настройки

Исследуем

— вердикты

Три правила честной игры

- 01 Всё измерено на GigaChat — модели Сбера
- 02 Рубли — по публичным тарифам для физлиц
- 03 Проценты разных опытов не складываем

Модель	Пакет	Цена пакета	₽ за 1 млн токенов
Lite	20 млн токенов	1 300 ₽	65
Pro	3 млн токенов	1 500 ₽	500
Max	3 млн токенов	1 950 ₽	650

Тарифы GigaChat для физлиц, с НДС, на 7 октября 2026 · цена за 1 млн рассчитана из цены пакета



MultiTester: из требований — в тест-кейсы



Требования



Мультитестер



Тест-кейсы

Пилот: ≈ 100 команд · 600 человек

72 % тест-кейсов приняты тестировщиками с первого раза

итоги пилота

Сегодня считаем только подготовку кейсов



Доклад на SQA DAYS

Внутри — конвейер агентов



1 запрос = 11 вызовов Max

Файлы читает и сохраняет обычный код

Попросил
тест-кейсы.

Получил
Совещание агентов.

За свой счёт.



Одно требование: кнопка «Сбросить фильтры»

Имя: Анна

Статус: Новая

Сбросить фильтры

№	Заявитель	Статус
101	Анна	Новая

Страница 1 · сортировка по номеру

- 1 Активна, если задан хотя бы один фильтр
- 2 Сброс: имя пусто, статус «Все», страница 1
- 3 Сортировка и права — как были

Чего ждём от теста

ДО

Анна · Новая

101 Анна — Новая



ПОСЛЕ СБРОСА

Имя пустое · Статус «Все»

101 Анна — Новая

102 Борис — В работе

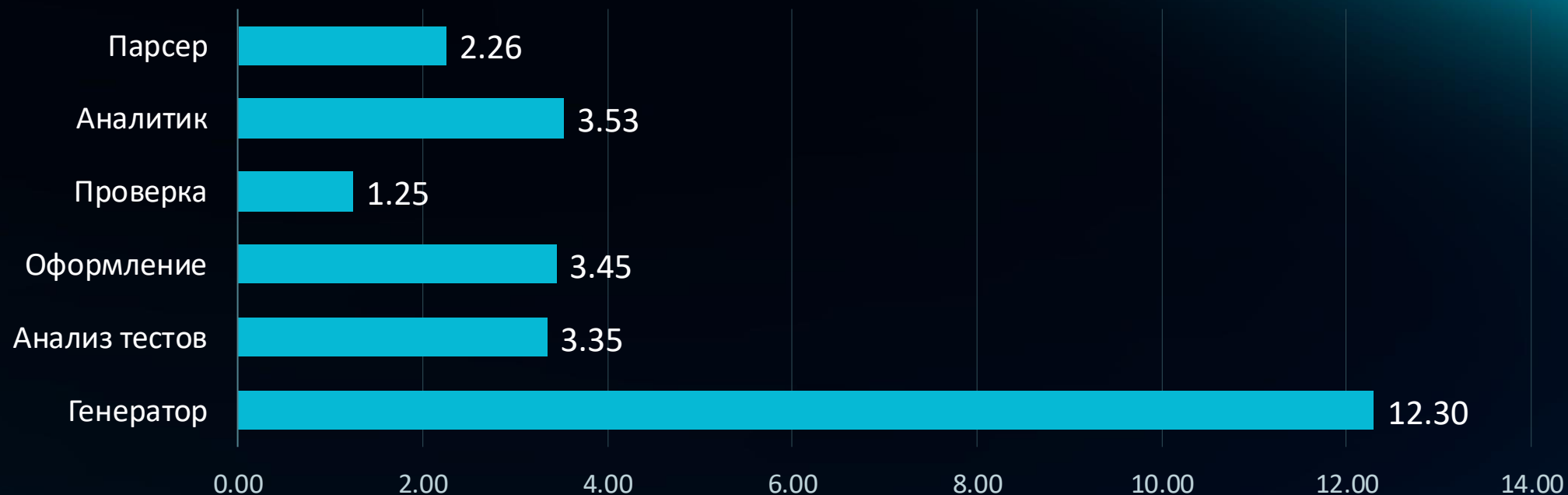
103 Анна — Завершена

Вот наш критерий приёмки

Счёт: первый проход

26,14 ₽

11 вызовов Max · 18 черновиков



Генератор — только половина счёта

Мелочь × поток



≈ 7,8 тыс. ₽

в день

≈ 150 тыс. ₽

в месяц

до 1,9 млн ₽

в год

Оценка по нагрузке пилота: до 300 задач в день

Настоящий счёт

26,14 ₽

первый проход:
18 черновиков

+ 9,93 ₽

исправление всех
непринятых кейсов

+ 7,20 ₽

дополнительная точечная
правка 6 кейсов

43,27 ₽

ИТОГО

Комплект тест-кейсов принят не полностью

Как уменьшить счёт

- 1 Не платить дважды
- 2 Не нести лишнее
- 3 Не переплачивать за модель

Работает

Требует настройки

Исследуем

Сначала — журнал

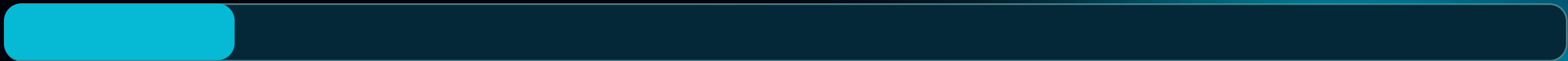


ID задачи	агент	модель	вход	кеш	ответ	исход
T-042	генератор	Max	9 111	3	9 817	не принят

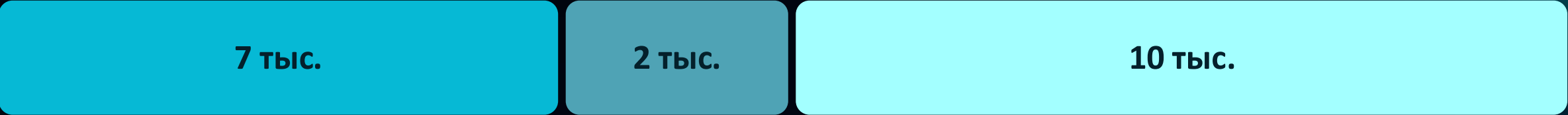
пример строки

Каждый вызов — строка. Повтор — тоже строка.

Один вызов генератора: что лежит на столе



19 тыс. из 128 тыс. токенов



Инструкции

правила + материал сборки

Запрос

требование + план проверок
от предыдущих агентов

Ответ

18 черновиков кейсов

Платим за всё, что на столе

Лучший вызов — тот, которого не было

Отдельный опыт: настоящий документ требований с таблицами

Код переносит
текст и таблицы



Модель размечает
разделы



Код сверяет
каждую ячейку

254 из 254

ячеек таблиц исходника — на месте

Работает

Убрали модель — добавили проверку

Четыре места для кеша

01

Сервис

одинаковое
начало запроса

02

Приложение

готовый
разбор

03

Комплект

готовый
результат

04

По смыслу

похожий
запрос

От простого — к рискованному

Серверный кеш

Парсер Мах · тот же документ · два одинаковых запроса · вход 5 862 токена

Первый запрос



Повтор



■ платно ■ из кеша

-73 %

Работает

Неизменное — наверх, меняющееся — вниз

Запрос, при котором кеш срабатывает

Так кеш не срабатывает

X-Session-ID: новый uuid

каждый раз новая сессия

system: «12.10.2026 14:03:27.

Ты — парсер. Правила...»

дата в первой строке — начало каждый раз другое

user: документ

Так кеш срабатывает

X-Session-ID: doc-4711

одна сессия на документ

system: «Ты — парсер. Правила...»

неизменное — наверху

user: документ, версия 79

user: задача · 12.10.2026 14:03:27

меняющееся — внизу

Работает

Проверка: в ответе `precached_prompt_tokens > 0`

Кеш готового разбора



5,75 ₽

разбираем заново

→ **1,78 ₽**

читаем из кеша

–69 %

Первый разбор оплачиваем: ≈ 5,7 ₽

Работает на повторях

Окупается со второго обращения

Кеш готового комплекта

20 из 20

совпали

0

вызовов модели

0,15 мс

чтение

Где пригодится

Повтор после сбоя

Одно требование —
несколько команд

Пакет: изменилось
1 из 100

18 кейсов: 12 из первого прохода + 6 доработаны отдельно

Работает

Повтор не стоит ни одного токена

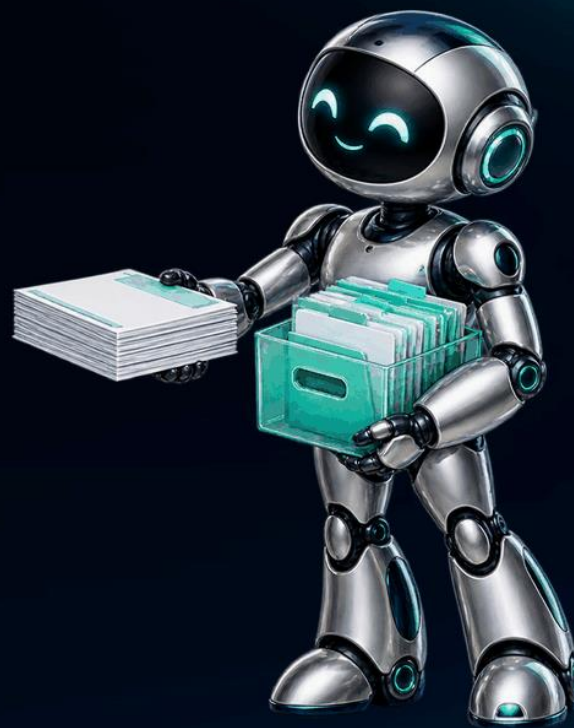
Модель

«Давайте я ещё раз
подумаю».



Кеш

«Мы уже за это
платили».



Кеш по смыслу

«Сохранить сортировку»

«Сбросить сортировку»

0

ложных попаданий

5 из 5

перефразировок не узнал

Поиск по ключевым словам + по векторам · как сделано — в памяти

Исследуем

Как точный кеш, только сложнее

А кеш ещё свежий?

Ключ кеша

SHA-256
требования

версия
инструкций

модель +
параметры

права
доступа

TTL 600 с: срок истёк — запись не отдаём

Устаревший ответ приходит очень быстро.

Это его единственное достоинство.

Глава 1. Не платить дважды

Работает

Код вместо модели

Работает

Серверный кеш

Работает

Кеш разбора

Работает

Кеш комплекта

Исследуем

Кеш по смыслу

Не думать меньше — не думать дважды

2

Не нести лишнее



Отбираем фрагменты требований вручную

Другой документ: сохранение данных организации · инструкции те же

7,8 → 4,6 тыс.

токенов на входе:

весь документ → нужные фрагменты

6,46 → 4,60 ₽

цена генерации

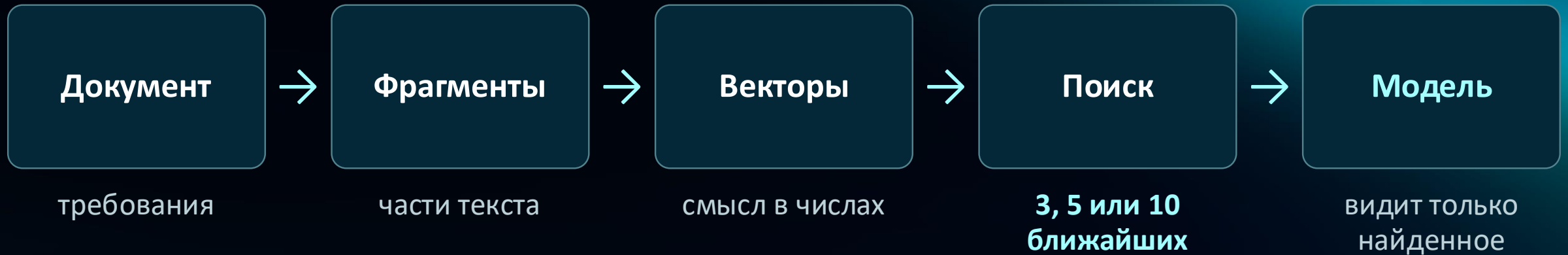
–29 %

Комплект отклонён

Требуется настройка

Настроить: все правила на месте

Как работает поиск



RAG — ответ модели по найденным фрагментам

Проверяем поиск до генерации

10 контрольных вопросов · вопрос засчитан, если найдены все нужные фрагменты

вернул 3 фрагмента

3 из 10

вернул 5 фрагментов

4 из 10

вернул 10 фрагментов

5 из 10

вопросов: всё нужное нашлось



Мастер-класс А. Волкова:
AI-ассистент с RAG

Исследуем

Сначала чиним поиск

Сжатие текста: LLMLingua-2

3 тыс. → 1,5 тыс. токенов

В требовании 53 поля, у каждого — своё правило
После сжатия все 53 названия на месте

status: не обновлять

→ **status: не**



LLMLingua-2

Все 53 названия
полей сохранили.

А что с ними делать...

Это мы сократили.

Требует настройки



Упаковка JSON: без потерь

токенов: с отступами → без отступов → со словарём



Работает

Упаковали и проверили

Пять вызовов → один

Аналитик: 5 субагентов, каждому — инструкции + весь документ

19,13 ₽

5 вызовов субагентов

→ **4,63 ₽**

1 общий вызов

−76 %

Потеряно правило замены состава. Результат отклонён

Требуется настройка

Настроить: каждое правило дошло

Глава 2. Не нести лишнее

Работает

Упаковка JSON

Требуется настройки

Отбор фрагментов вручную

Требуется настройки

Сжатие текста

Требуется настройки

Объединение вызовов

Исследуем

Поиск фрагментов (RAG)

Меньше токенов ≠ экономия. Сначала настройка

3

Не переплачивать за модель



Каскад: сначала дешёвая, потом дорогая

Lite 0,47 ₽



Max 4,63 ₽

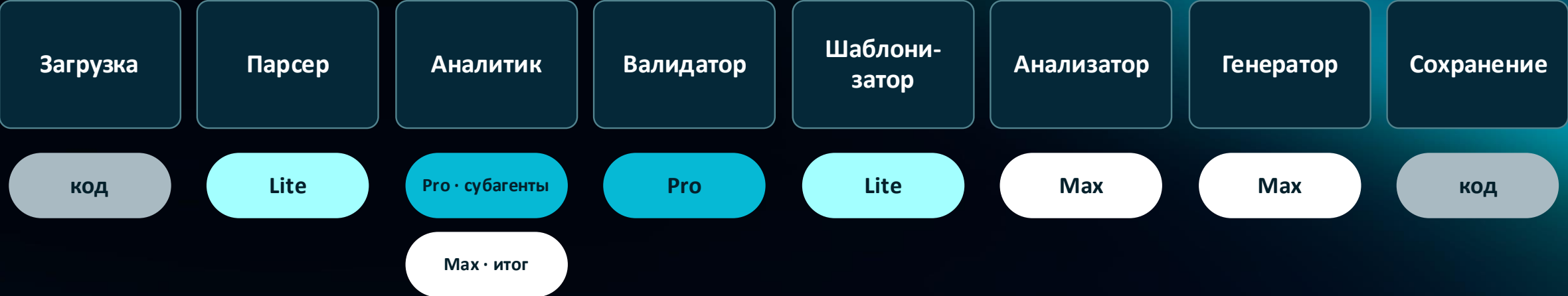
= 5,11 ₽

Оба результата отклонены

Требует настройки

Настроить: когда хватает дешёвой

Модель — по роли



Работает

Не одна модель на всё — выбор под задачу агента

Исследуем

Batch

пачкой на ночь

Своя модель

платим за железо

LoRA

дообучение под задачу

Открытые веса GigaChat: Lightning — 10 млрд параметров, 3.5 Ultra — 432 млрд

Исследуем

Памятки

Карта результатов

Работает

Код вместо модели

Серверный кеш

Кеш разбора

Кеш комплекта

Упаковка JSON

Модель по роли

Требует настройки

Отбор фрагментов вручную

Сжатие текста

Объединение вызовов

Каскад моделей

Исследуем

Поиск фрагментов (RAG)

Кеш по смыслу

Batch

Своя модель

LoRA

Не повторять — сразу. Давать меньше — после настройки.

А если у вас не GigaChat?

Памятки: техники и другие LLM

GPT · Claude · Gemini · DeepSeek · Qwen

Переносится метод.

Не проценты.

Эксперимент на понедельник

1 Определить результат

2 Измерить весь путь

3 Найти самое дорогое

4 Изменить одно

5 Сравнить цену и качество

6 Проверить на новом

Одно требование. Один журнал. Одно изменение.

Результат принят.

Качество сохранено.

Полная стоимость меньше.

Памятки: техники и другие LLM

Считайте не токены.

**Считайте цену результата,
которому можно доверять.**

Спасибо!

Евгений Гребенщиков

@COOLLight
t.me/ITvsAI



Памятки: техники и код,
другие LLM

