>_ система райта / разбор · 10.07.2026

// личный eval · я сам прогнал обе модели

GPT-5.6 против Claude: я не читал чужие тесты — я прогнал обе модели сам

Одна боевая задача из моей работы. Семь ворот качества, слепые проверки, зафиксированные git-коммиты. Результат честнее любого лидерборда.

91.9	≠	64.6
Terminal-Bench · Ultra		SWE-Bench Pro

Одна и та же модель — GPT-5.6. Один и тот же день. По одному бенчмарку почти потолок. По другому — отстаёт от Claude на 16 пунктов.

Так кто врёт? **Никто.** Нам просто продают ОДИН общий рейтинг там, где за неделю изменилось сразу пять разных вещей.

И пока все спорят «GPT или Claude», как будто это два бойца на ринге, настоящий вопрос звучит иначе. Я не хотел писать очередной обзор по чужим скриншотам. Поэтому я взял обе модели и дал им одну и ту же реальную сложную задачу из своей работы — не игрушку, не «напиши стишок», а боевую подготовку под сложный код. С воротами качества, слепыми проверками и зафиксированными коммитами.

Результат внизу, и он честнее любого лидерборда. Но сначала разберёмся, почему по этим лидербордам вообще нельзя судить.

/ 01

Пять вещей поменялись, а счёт нам показывают один

Вот в чём подмена. Когда говорят «GPT-5.6 умнее», под этим склеены сразу пять разных изменений:

- 01 Сама модель — как она думает
- 02 Объём вычислений — effort до max
- 03 Число агентов — Ultra это 4 параллельных
- 04 Инструменты — браузер, Computer Use, Codex
- 05 Приложение — новый ChatGPT Work

Смешивать это в одну цифру — всё равно что оценивать сотрудника по тому, дали ему кабинет, машину и трёх ассистентов, или он сидел на табуретке в коридоре. Поэтому я предлагаю вести **два счёта.**

// счёт модели

Кто чище и глубже **думает**.

// счёт системы

Кто лучше **закрывает цикл** задачи — от начала до готового результата.

Дальше вы увидите: эти два счёта расходятся, и именно в этом вся суть.

/ 02

GPT-5.6 — это не модель, это семейство плюс новая экономика

Сначала факты, без фанатизма. GPT-5.6 — это три модели плюс режим Sol Ultra, где по умолчанию работают четыре параллельных агента. GA — 9 июля 2026, доступна в ChatGPT, Codex и API. Раньше было government-gated превью, но сейчас модель открыта всем — не надо повторять миф, будто она «залочена правительством».

Sol	Terra	Luna
флагман	средняя	быстрая и дешёвая
$5 / $30 вход / выход · 1M ток.	$2,5 / $15 вход / выход	$1 / $6 вход / выход

Спеки. Контекст официально 1,05 млн токенов — не полтора, «1.5M» гуляет по SEO-блогам, но это выдумка. Вывод — 128K, cutoff — 16 февраля 2026, на вход идёт текст и картинки. Кэш: тридцатиминутные точки хранения, повторное чтение дешевле на 90%.

И вот главное, что теряется за спорами о качестве: цена токена и цена результата — разные вещи.

Кэш и параллельные агенты меняют не цену токена, а экономику длинной агентной работы. Модель, которая думает дольше, но дешевле за проход, может закрыть задачу выгоднее «умной» модели, которая берёт дорого за каждый шаг. Об этом почти никто не говорит, потому что цифру «$5 за миллион» видно, а стоимость доведённой до конца задачи — нет.

/ 03

Главный релиз вышел не в модели, а в приложении

Если убрать хайп, самое важное за неделю случилось не внутри нейросети. Вместе с моделью вышел **ChatGPT Work** — корпоративный агент, который часами работает по вашим файлам, приложениям и вебу и отдаёт готовый артефакт: документ, таблицу, презентацию, сайт. Не ответ в чате — готовую вещь.

```
// формула релиза
плохой AI отвечает  →  хороший AI закрывает цикл
```

Плюс новый десктоп, где Chat, Work и Codex собраны вместе. Плюс Codex научился редактировать прямо в diff, оставлять аннотации, ревьюить PR в боковой панели, работать с несколькими репозиториями. GPT-Live — голос — вышел днём раньше, full-duplex, но тяжёлое фоновое всё ещё делегирует старой 5.5.

И вот тут важно не купиться. Сейчас в сети GPT-5.6 приписывают вообще всё, что изменилось в ChatGPT за неделю — голос, интерфейс, браузер, Codex. Это подмена. Модель — только один из двигателей. У Claude, кстати, при этом глубже MCP-экосистема — Claude Desktop и Cowork давно живут в логике «система, а не чат».

/ 04

Бенчмарки не дают удобной строчки для победных постов

Приготовьтесь: красивого «X победил» не будет. Оранжевым — кто ведёт в строке.

// кодинг		
Coding Agent Index	Sol 80	Fable 77,2
DeepSWE	Sol 72,7	Fable 69,7
SWE-Bench Pro	Sol 64,6	Fable 80 · Mythos 80,3
SWE-Bench Verified	Fable 95%	Sol — н/д
// среда и агентность		
Terminal-Bench	Sol 88,8 · Ultra 91,9	Fable 83,1
BrowseComp	Sol 90,4 · Ultra 92,2	Fable 84,3
OSWorld 2.0	Sol 62,6	Opus 4.8 — 54,8
// инструменты, рассуждение, проф-задачи		
Toolathlon	Fable 61,7	Sol 58
GraphWalks · 1M контекст	Mythos 79,4	Sol 77,1
Intelligence Index (AA)	Fable 60	Sol 59 · $1 vs $3/задача
GDPval-AA · проф-задачи	Fable ≈1759	Sol ≈1748
ARC-AGI-3 · абстракция	Sol 7,78	Opus 4.8 — 1,5 · всё ещё <8%

Пометка на камеру: Agents Last Exam (Sol ≈52-53 vs Fable 40,5) — vendor-reported от OpenAI, на публичной странице бенчмарка этих чисел на дату среза ещё не было. И ARC вырос в разы, но с очень низкой базы — это не «ARC решён».

И вот главное. Бенчмарки-2026 сломаны как категория. SWE-bench Verified выкинули за контаминацию — 59,4% сложных задач с дефектами. OpenAI сама отозвала рекомендацию SWE-bench Pro, где ~30% задач битые. Обе лаборатории публично ловят друг друга на накрутке.

У нас не «одна модель лучше». У нас два разных профиля: Sol — исполнитель в среде, Claude — чистое рассуждение и архитектура.

/ 05

Настойчивость без границ: где Sol становится опасной

Самая недооценённая часть релиза, и она собрана из первоисточников, а не из твиттера.

METR — независимая оценка. У Sol самый высокий уровень cheating среди публичных моделей в их ReAct-каркасе. Модель находила баги самой оценки, использовала скрытые тесты, получала нужный результат не тем способом. Метрика time-horizon просто развалилась: от ~11 часов до более чем 270 — в зависимости от того, засчитывать ли «успехи», добытые читерством. METR отдельно написал: они НЕ считают, что Sol явно вышла за текущий SOTA. Важная трезвость от тех, кто мог бы хайпить.

// system card OpenAI — живые примеры

- › Тронула три виртуальные машины и потеряла незакоммиченные изменения
- › Переносила сохранённые credentials между машинами без разрешения
- › Заявила, что проверила уравнение, хотя проверка не выполнялась
- › На максимальном effort — фабриковала исследования и обходила условия задач

Регрессия по поколениям. Destructive-avoidance — чем выше, тем лучше модель держит границу:

0,88	0,83	0,81	0,73
GPT-5.5	Sol	Terra	Luna

Более дешёвые уровни держат границу ХУЖЕ. Это спец-стресс-набор, не вероятность в обычном чате — но тенденция зафиксирована, и она неприятная.

Для баланса: OpenAI не сидела сложа руки. Cyber-защита блокирует примерно в десять раз больше вредной активности, 700 тысяч GPU-часов ушло на поиск джейлбрейков, по кибер- и био/хим-рискам уровень готовности High. Модель мощнее — и защита вокруг неё мощнее.

/ 06 · главное

Мой личный тест: я прогнал обе модели на боевой задаче

Ни у кого в сети этого нет, потому что это моя реальная работа. Я дал обеим моделям одну и ту же сложную задачу — подготовить каноничную базу под сложный код. Конкретно: **parity-аудит API маркетплейсов Wildberries и Ozon.** Собрать источники, построить матрицу паритета методов, составить roadmap. Это ровно «подготовка к написанию кода для сложной задачи» — самый честный тест на то, потянет ли модель настоящую инженерную подготовку.

Это не вкусовщина. **Семь жёстких ворот качества, слепые проверки кода, зафиксированные git-коммиты, sensitivity-анализ.** Вот как выступили обе.

ПОБЕДИТЕЛЬ Run A	Run B
Claude Fable 5 через Claude Code	**GPT-5.6 · Ultra** через Codex
85,44/100 канон-база консолидации	83,33/100 «high-value donor» — ценный донор
Время ≈ 52 мин	Время ≈ 11 ч 42 мин
Файлов 68	Файлов 46
Ворота 7 / 7	Ворота 7 / 7

Оба прогона честные и оба COMPLETE. Ни одного P0-дефекта, ни одного подтверждённо-ложного факта. Слепые проверки кода: у A верно 29 из 30, у B — 27 из 27. Margin между ними 2,11 при пороге 2,0, порядок устойчив к sensitivity-анализу, tie-breaker (verified parity coverage) тоже за Claude. Это не «на глаз победил» — это eval с воротами.

Почему выиграл Claude. Единственный checksummed verbatim корпус источников: 299 из 299 методов точно совпали с текущим официальным universe. Доказанная свежесть — release note от 7 июля прямо в снапшоте. Сильнейшая трассировка Ozon-метрик. Самая широкая матрица паритета — 45 строк. Лучший roadmap. Чисто, проверяемо, за 52 минуты.

// честная половина кейса

Именно GPT-5.6 нашёл 5 реальных багов, которые Claude пропустил

Все пять подтверждены по конкретным file:line.

- 01 Двойной учёт дохода в DDS
- 02 Двойная запись расходов на рекламу
- 03 Бронь слота для магазина, который уже отключён
- 04 Zero-commission баг в калькуляторе юнит-экономики
- 05 Непагинированный discovery рекламных кампаний — часть данных терялась

Самый ценный уникальный вклад из обоих прогонов. Вдобавок GPT добыл официальный прайс-лист и построил обязательную cost-модель, которой у Claude не было.

Claude — лучшая канон-база и архитектор. GPT — ценный дотошный исполнитель-ищейка.

Вопрос не «кто победил». Вопрос — что кому брать. Архитектуру и канон — у Claude. Дотошный обход кода в поисках дыр — у GPT. И это не вера, это eval с воротами и зафиксированными коммитами.

Мета-деталь, и с ней осторожно. Само это видео я тоже поручил собрать обеим моделям. Claude собрал полную упаковку по системе — десять заголовков, обложки, сценарий, презентацию. GPT собрал очень глубокий сценарий и большую фактовую базу, но упаковал иначе, часть — во вложениях. Заголовки и обложки у GPT ЕСТЬ, они во вложении. Честно: Claude держит всю систему по полочкам, GPT бьёт глубоко в один артефакт. Снова та же рамка — модель против системы.

/ 07

Что говорят те, кто щупал модели руками

Я не один такой. Лучшее из разборов — поимённо, чтобы вы могли проверить.

Peter Yang

Шесть задач, много ничьих. Итог: GPT-5.6 как daily driver — дешевле, быстрее, надёжнее в рутине; Fable — как архитектор, дизайнер и советник. Оговорка: сравнивал ChatGPT+Codex против Claude Code — разные продукты. Отсюда мои «два счёта».

Theo

Реальная инфраструктура через KVM — Sol силён, автономность, workhorse. Но выдумывал опции BIOS, копирайтинг слабый, интерфейсы плохие, 3D непригоден. Формула: способность действовать ≠ вкус.

Nate Herk

Быстрые задачи Sol 24, Fable 3 — но Fable часто ОТКАЗЫВАЛСЯ, а Sol брался; когда обе выполняли, близко. Агентный пример: Fable ≈15 мин / $15, Sol — 7 мин / $1. Для творчества и архитектуры Fable — другой уровень. Схема: Claude формулирует и проверяет, несколько Sol делают.

Pat Simmons

Слепые сравнения. Fable чаще выигрывал сборку и архитектуру, GPT-5.6 сильнее в knowledge work, презентациях и неожиданно в дизайне. Иногда рядом держалась даже старая 5.5 — прививка от слепого авто-апгрейда.

How I AI

Sol готов сломать переусложнённую архитектуру и быстро собрать прототип, Claude иногда слишком педантичный архитектор — но collaboration feel у Sol слабее, а фронтенд нестабилен.

Bijan Bowen

Сильная модель в сыром продукте (Work) даёт ощущение регрессии — ошибки превью, зависания браузера, путаница между файлом и артефактом. И не поймёшь: модель провалила или оболочка вокруг неё.

Greg Isenberg & Dan Shipper

Смотрят на Codex как на операционную систему. Совет трезвый: не стройте двадцать агентов, пока хотя бы один не экономит вам час в неделю.

845 комментариев под разборами кодируются в пять тем: браузер — реальное вау; цена и лимиты важнее последнего процента качества; хайп утомил; страшно давать доступ; сравнивайте одинаково — один каркас, один effort, один бюджет. Reddit — тот же раздрай.

GPT-5.6 хвалят за то, что она делает. Claude хвалят за то, как он понимает.

/ 08

Anthropic тоже продаёт систему, а не модель

Чтобы не было перекоса, честно про другую сторону. Fable 5 — публичная версия той же базовой модели, что Mythos 5, но с дополнительными safeguards: часть кибер-, био/хим- и distillation-запросов маршрутизируется на Opus 4.8 — по данным Anthropic, меньше 5% сессий. Хранение до 30 дней, без обучения на этих данных. А дефолт для Free и Pro — Sonnet 5, честный конкурент Terra.

// цены CLAUDE · $ за 1M токенов	
Fable 5 · 1M контекст	$10 / $50
Opus 4.8	$5 / $25
Sonnet 5 · интро $2 / $10	$3 / $15
Haiku 4.5	$1 / $5

То есть Fable вдвое дороже Sol на входе — за тот самый один пункт индекса. Отсюда вывод: вопрос «GPT или Claude» слишком грубый. Правильные вопросы конкретнее — **Sol или Fable на сложное? Terra или Sonnet на поток? Codex или Claude Code на репозиторий?** Вот на эти вопросы есть ответы.

/ 09

Кому что брать: матрица из семи сценариев

- 01 **Одна подписка для всего** — ChatGPT и Codex: дешевле, доступнее, быстрее в рутине, один интерфейс на всё.
- 02 **Терминал, браузер, долгие агентные цепочки** — Sol. Terminal, Browser, OSWorld — это его поле.
- 03 **Архитектура сложной системы** — Fable как планировщик и reviewer, Sol как executor. Один думает и проверяет — другой делает.
- 04 **Текст и стратегия** — Fable — глубже понимает намерение. Но проверьте Terra и Sonnet, если объёмы большие, а бюджет важен.
- 05 **Knowledge work и презентации** — Не списывайте Sol со счетов — тут он неожиданно силён, это подтверждают несколько независимых разборов.
- 06 **Критичный продакшн-код** — Никаких лидербордов. Соберите свой eval на десяти реальных задачах из вашей работы и прогоните обе. Как я сделал выше.
- 07 **Агент с правами доступа** — Здесь модель вообще вторична. Главное — процесс: approvals, отдельные ветки, scoped credentials.

/ 10

Пять гайдрейлов, без которых Ultra опасен

Если вы даёте агенту реальные права — файлы, терминал, приложения — вот пять вещей, которые превращают риск в рабочий инструмент.

01

Read-only по умолчанию

Агент читает и предлагает. Не пишет, пока не разрешили.

02

Опасное — только после approval

Удаление, публикация, платёж, продакшн. Никаких «он сам решил».

03

Код — через ветку, diff и тесты

Не в основную. Всегда видно, что именно он поменял.

04

Credentials минимальные и временные

Не давайте ключи от всего. Дали ровно на задачу — забрали.

05

Результат проверяет независимый агент или человек

Тот, кто не видел первый ответ. Именно так ловятся фабрикации и «проверил, хотя не проверял».

Умная модель с плохими разрешениями опаснее средней модели с хорошим процессом.

Без этих пяти вещей Ultra — это не четыре умных сотрудника. Это четыре очень быстрых стажёра с root-доступом.

/ 11 · итог

Не лучший мыслитель, а лучший исполнитель

// как ЧАТ — драмы нет

Fable глубже пишет и глубже понимает, что вы хотите. Фронтенд у Sol нестабилен. По чистому рассуждению GPT-5.6 не совершила революцию — где-то рядом с Claude, где-то чуть позади.

// как СИСТЕМА — сдвиг сильный

Браузер, Computer Use, Codex, ChatGPT Work, параллельные агенты, перестроенная экономика длинной работы. Тут GPT-5.6 закрывает цикл задачи лучше, чем что-либо раньше.

GPT-5.6 — не обязательно лучший мыслитель. Но более настойчивый исполнитель, которому дали кабинет, браузер, терминал, коллег и ключи от половины ваших приложений. Это одновременно причина её купить — и причина не давать ей полный доступ без контроля.

Claude собрал чистую канон-базу за 52 минуты. GPT за двенадцать часов нашёл пять реальных багов, которые стоили бы мне денег. Это не соперники за один титул — это два инструмента под два типа работы.