🤖 Сравнение Claude Opus 4.8 и GPT-5.5 на продакшн-коде

🚀 Сравнение Claude Opus 4.8 и GPT-5.5: реальный тест на продакшн-код

Ключевые тезисы:

- Вышла новая модель **Claude Opus 4.8** с улучшенными бенчмарками и новыми функциями (мультиагентный режим, workflow).
- Практический тест на реальной бизнес-задаче (адаптация AI-бота для агентства недвижимости) показал противоречивые результаты.
- **GPT-5.5** точнее следует промту и создаёт более функциональную админ-панель, но требует больше токенов (лимиты урезаны).
- **Opus 4.8** справился с задачей за один подход, но сильно отошёл от исходного шаблона, ухудшив некоторые функции.
- Идеальный стек на май 2026 года — комбинация обеих моделей, но, вероятно, уже на тарифах по $100.

📊 Теоретическая часть: что нового в Opus 4.8?

- **Цены:** Остались прежними: $5 за млн входных и $25 за млн выходных токенов. Появился **Fast Mode** (в 3 раза дешевле, чем у 4.7).
- **Режимы работы:** Помимо уровня *High*, появились *Max* и *Ultra* для максимальной производительности.

- **Бенчмарки (основные):**
  - **Terminal Bench 2.1 (работа в терминале):** Opus 4.8 (74.6) улучшил результат, но всё ещё позади GPT-5.5 (78.2).
  - **SWE Bench Pro (программирование, фиксы багов):** Opus 4.8 (69) против 4.7 (64) — значительный прогресс.
  - **Computer Use (работа через браузер):** Opus 4.8 (83.4%) — **лидер** в категории.
  - **Super Agent Benchmark (длинные бизнес-цепочки):** Opus 4.8 на уровне GPT-5.5.
- 🔥 **Ключевое улучшение: Честность модели.** Заявлено, что Opus 4.8 **в 4 раза реже пропускает баги** и честнее признаётся, если не справился с задачей, что экономит часы дебага.
- **Новые функции API:** Системные инструкции теперь можно обновлять в процессе диалога (как у OpenAI), что удобно для агентских цепочек.

⚙️ Новая фича: мультиагентный режим (Workflow)

- **Суть:** Раньше Claude работал как один агент. Теперь для сложных задач он **сам пишет скрипт на JavaScript**, разбивает задачу на части и запускает до **16 субагентов параллельно** (до 1000 за сессию).
- **Активация:** Добавить слово `workflow` в промт или команду `effort Ultra`.
- **Сравнение с Cursor (Codex):** У Cursor максимум 6 субагентов, нет оркестрации workflow, нельзя сохранять и переиспользовать сценарии.

🧪 Практический тест: AI-бот для недвижимости

Задача: Адаптировать готовый продакшн-стартеркит (7600 строк кода, 47 файлов) под агентство недвижимости, не трогая ядро.
Методология: Один промт для обеих моделей. Тестирование функциональности и перекрёстный code review по 8 критериям.

🔎 **Результаты выполнения задачи**

- **GPT-5.5 (Codex):**
  - ✅ **Точно следовал промту**, минимально изменив стартеркит.
  - ✅ Админ-панель и виджет получились **функциональными и логичными** (работают блокировки, перехват диалогов, выгрузка лидов).
  - ❌ **Потребил все лимиты** подписки за $20, пришлось докупать кредиты. Лимиты ощутимо урезаны.
  - ⚠️ Мелкий баг в статистике (некорректный вывод числа "горячих лидов").
- **Claude Opus 4.8:**
  - ✅ Справился **за один подход**, не запрашивая доп. действий.
  - ✅ **Лучший визуал и UX:** создал целый лендинг, добавил подтверждения для опасных действий (например, блокировки).
  - ❌ **Сильно отошёл от шаблона**, ухудшив ключевую функциональность: неработающая блокировка пользователя, нельзя открыть загруженные документы в базе знаний.
  - ❌ **Медленнее** в генерации ответов в готовом виджете.

📝 Результаты перекрёстного Code Review

Критерий	Opus 4.8 оценил свой код	GPT-5.5 оценил код Opus	GPT-5.5 оценил свой код	Opus 4.8 оценил код GPT
Итоговый балл	**60/80**	**45/80**	**50/80**	**56/80**
Ключевой вывод	Крепкая, безопасная основа. Минусы: падающие тесты, баг в конфигурации.	Код "сырой", проблемы с продакшн-готовностью и тестами.	Рабочий, но "сшит на скорую руку" кандидат в продакшн.	Реальный работающий продукт на крепком ядре. Чистый код.

Вывод по оценкам: GPT-5.5 остаётся **более строгим и критичным** оценщиком как своего, так и чужого кода. Opus 4.8 оценивает **более лояльно**.

🎯 Идеальный стек инструментов на май 2026 года

- **Прошлая рекомендация** (ChatGPT $20 + Claude $100) **устарела** из-за урезания лимитов у OpenAI.
- **Текущая ситуация:** Для серьёзной разработки, скорее всего, понадобятся тарифы по **$100 в обеих экосистемах**.
- **Claude Code ($100):** Силён в **мультиагентном режиме (workflow)** и имеет **Code Review Ultra**. Подходит для крупных проектов с большими лимитами.
- **ChatGPT/Codex ($100):** Точнее следует инструкциям, даёт более жёсткий и полезный code review.
- 🏆 **Итоговый вердикт: Нет однозначного победителя.** Лучшая стратегия — использовать **обе модели в связке**, переключаясь между ними в зависимости от задачи.
  - **Opus 4.8** — для быстрого прототипирования, сложных параллельных задач и workflow.
  - **GPT-5.5** — для точной реализации по ТЗ, строгого код-ревью и задач, требующих дословного следования промту.
