🤖 Сравнение китайских AI-моделей для генерации кода

🏆 Сравнение китайских AI-моделей для генерации продакшен-кода

Ключевые тезисы

- Китайские модели (GLM 5.1, Qwen 3.6, Kimi 2.6, DeepSeek-V4 Pro) сравнили по способности адаптировать реальный боевой стартер-кит (7600 строк кода).
- Критерии оценки: безопасность, модульность, тесты, документация, обработка ошибок, готовность к продакшену, качество кода, функциональность.
- Лучшей по сочетанию скорости, стоимости и качества реализации стала **GLM 5.1**.
- **GPT-4o** и **Claude 3.5 Sonnet** провели перекрёстный код-ревью, выявив тенденцию Claude к завышению оценок.

🎯 Методология эксперимента

Цель: Проверить, могут ли китайские модели создать готовый к деплою код для AI-продавца в новой нише (мебельный салон) на основе существующего продвинутого стартер-кита.

Инструменты:

- **Стартер-кит:** Продакшен-код AI-продавца (7600 строк, 47 файлов, тесты, админ-панель, защита от инъекций).
- **Open CMD:** Open-source инструмент для анализа кода.
- **Open Router:** Агрегатор моделей с оплатой за использование.

Испытуемые модели:

1. **DeepSeek-V4 Pro:** 600B параметров, контекст 1M токенов.

2. **Kimi 2.6 (Moonshot AI):** Триллион параметров, нативная агентная модель.
3. **Qwen 3.6 (Alibaba):** 27B параметров, открытая лицензия Apache 2.0.
4. **GLM 5.1:** 754B параметров, может автономно работать над задачей до 8 часов.

Процесс: Одинаковый промпт без подсказок → генерация кода → функциональное тестирование (виджет, админка, защита) → самооценка модели → перекрёстный код-ревью от GPT-4o и Claude 3.5 Sonnet.

🔍 Результаты функционального тестирования

GLM 5.1 ✅

- **Стоимость:** ~$4 (100K токенов).
- **Результат:** Лучший. Виджет и админка работают корректно. Защита от промпт-инъекций срабатывает. Админ-панель функциональна (диалоги, лиды, база знаний, настройки с выбором моделей).
- **Недочёты:** Нельзя перейти из лидов в диалог.

Qwen 3.6 ⚠️

- **Стоимость:** ~$10 (244K токенов).
- **Результат:** Виджет не работает (проблема с подключением API). Админка минималистична, но есть баги (перемешались диалоги с другим проектом, нет ключевых настроек SLA/блокировок).

Kimi 2.6 ⚠️

- **Стоимость:** ~$5 (170K токенов).
- **Результат:** Виджет работает, защита от инъекций есть. Админка сырая: диалоги нельзя нормально прокрутить, база знаний не удаляется, реализация хуже предыдущих.

DeepSeek-V4 Pro ❌

- **Стоимость:** ~$11.2 (200K токенов).

- **Результат:** Худший. Виджет не работает. Админ-панель не открывается (нерабочая кнопка входа). Деньги и время потрачены впустую.

📊 Результаты код-ревью (оценка из 80 баллов)

GLM 5.1:

- **Самооценка (GLM):** 45/80
- **Оценка GPT-4o:** 45.8/80
- **Оценка Claude 3.5 Sonnet:** 59/80
- **Вывод:** GPT-4o и модель объективны, Claude завысил оценку.

Qwen 3.6:

- **Самооценка (Qwen):** 51/80
- **Оценка GPT-4o:** 52/80
- **Оценка Claude 3.5 Sonnet:** 55/80
- **Вывод:** Оценки близки, Claude снова немного завысил.

Kimi 2.6:

- **Самооценка (Kimi):** 51/80
- **Оценка GPT-4o:** 53/80
- **Оценка Claude 3.5 Sonnet:** 58/80
- **Вывод:** Тенденция сохраняется.

DeepSeek-V4 Pro:

- **Самооценка (DeepSeek):** 49/80
- **Оценка GPT-4o:** 52/80
- **Оценка Claude 3.5 Sonnet:** 68/80
- **Вывод: Claude дал необъективно высокую оценку** нерабочему коду, что ставит под сомнение его полезность для объективного код-ревью.

💡 Ключевые выводы

1. **Китайские модели догнали западные по качеству (в бенчмарках и на практике)**, но остаются значительно дешевле.
2. **GLM 5.1 — явный победитель** в этом тесте: быстрая, дешёвая и качественная реализация с адекватной самооценкой.
3. **GPT-4o (Codex) показал себя как отличный инструмент для объективного код-ревью**, его оценки совпадали с самооценками моделей.
4. **Claude 3.5 Sonnet склонен к завышению оценок**, особенно для кода других моделей, что делает его менее надёжным для этой задачи.
5. **Идеальный стек:** Использовать китайские модели (вроде GLM 5.1) для генерации кода, а GPT-4o — для его объективной проверки.