ПРАКТИЧЕСКИЙ ГАЙД · ИЮНЬ 2026 · ОТЧЁТ №3

5–30x

Китайский стек для бизнеса

Как строить продукты на DeepSeek, Qwen, Kimi и GLM: рейтинговые карточки моделей, юнит-экономика, каталог из 30 юзкейсов, маршруты подключения, юридические риски и чеклист внедрения. Не обзор — рабочий справочник.

61%

токенов OpenRouter в феврале 2026 пришлось на китайские модели — годом ранее было меньше 2%

5–30x

настолько дешевле западных флагманов обходится инференс сопоставимого качества

$252

месяц работы кодинг-агента на DeepSeek против $22 500 за тот же объём на западном флагмане

СОДЕРЖАНИЕ

Что внутри

Методика оценок

Каждая модель оценена по пяти осям из 10 баллов: **код и агенты** (SWE-Bench Pro/Verified, LiveBench, длина автономных сессий), **цена** (стоимость за 1M токенов с учётом кэширования и наличия бюджетной линейки), **скорость** (токены/сек и латентность), **открытость** (лицензия весов и реалистичность self-hosting), **экосистема** (доступность через агрегаторы и облака, полнота линейки, эмбеддинги, документация). Оценки — редакционные: они сводят публичные бенчмарки и ценовые листы в одну шкалу, но не заменяют тест на ваших задачах. Цены и метрики актуальны на 11 июня 2026 г.; цены меняются часто — проверяйте перед расчётом бюджета.

РАЗДЕЛ 01

Главное за 90 секунд

Пока западные команды по привычке подключают трёх провайдеров, на рынке сложился второй стек — китайский: пять фронтир-лабораторий с открытыми весами, OpenAI-совместимыми API и ценой **в 5–30 раз ниже** при сопоставимом качестве в коде и агентных задачах. Это уже не «дешёвая альтернатива», а инфраструктура, на которой работают флагманские западные продукты.

> Разрыв с лучшими проприетарными моделями — около 3–6 баллов из 60 по Intelligence Index. Год назад было втрое больше. Цена при этом ниже на порядок.

Решётка выбора: с чего начать под вашу задачу

ЗАДАЧА	ПЕРВЫЙ КАНДИДАТ	ЗАПАСНОЙ
Массовый дешёвый трафик: чат-боты, классификация, суммаризация	**DeepSeek V4 Flash**	GLM-4.7-Flash (бесплатный)
Кодинг-агент с лучшим балансом цена/качество	**MiniMax M3**	DeepSeek V4 Flash
Долгие автономные агентные сессии, оркестрация суб-агентов	**Kimi K2.6**	GLM-5.1
Мультиязычный продукт, длинный контекст (1M), надёжный tool-calling	**Qwen 3.6 Plus**	DeepSeek V4 Flash
Self-hosting с максимально свободной лицензией	**GLM-5.1 (MIT)**	Qwen 3.6 (Apache 2.0)
Локальная разработка на одной видеокарте 24GB	**Qwen3.6-27B**	—
Данные не должны касаться серверов в КНР	**Self-host / AWS Bedrock**	Together / Fireworks

Главная стратегическая идея гайда: **роутинг вместо моногамии**. Команды, которые маршрутизируют запросы — дешёвая модель на массовые подзадачи, сильная на критические — получают и качество, и экономику. Все китайские API совместимы с OpenAI SDK: смена модели — это смена одной переменной baseURL, поэтому цена эксперимента близка к нулю.

РАЗДЕЛ 02

Кто уже строит на китайском стеке

Лучший аргумент в пользу production-готовности — не бенчмарки, а то, кто на этом уже зарабатывает. Список получился неожиданным: это флагманы американского рынка.

Cursor — Composer 2 построен на Kimi

Самый популярный AI-редактор кода раскрыл, что его фирменная модель Composer 2 базируется на открытой Kimi K2.5 от Moonshot AI. Пользователи заметили неладное раньше официального признания — в выводе кода периодически проскакивали китайские иероглифы. Это не интеграция «для галочки»: Composer — ядро продукта, на котором Cursor строит подписочную выручку.

Windsurf — SWE-1.5 на базе GLM

Второй из культовых кодинг-инструментов подтвердил, что его модель SWE-1.5 предоставлена китайской Zhipu AI — та сама ретвитнула новость с поздравлениями. Контекст добавляет иронии: годом ранее Anthropic отключила Windsurf от своих моделей во время сделки с OpenAI — урок о том, что зависимость от одного западного вендора тоже риск.

Groq, Vercel, Cerebras, Together

Основатель Social Capital Чамат Палихапития о миграции Groq на Kimi K2: «заметно производительнее и попросту намного дешевле», чем OpenAI и Anthropic. Vercel добавил GLM в свой AI Gateway, Cerebras продвигает GLM как основную модель на своих чипах, Together размещает Qwen-Coder. Инфраструктурный слой американского AI-рынка голосует токенами.

Обратная сторона: регуляторное внимание

В апреле 2026-го комитеты Конгресса США направили письма Cursor и Airbnb с вопросами об использовании китайских моделей. Запретов нет, но направление давления понятно: если ваш рынок — госсектор или чувствительные отрасли США, держите это в матрице рисков (раздел 11).

61%

токенов OpenRouter — китайские модели (февраль 2026)

7 из 10

самых используемых моделей платформы — китайские (май 2026, Bloomberg)

25 трлн

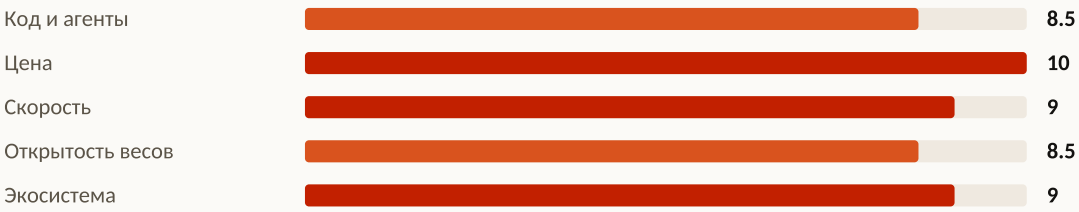
токенов в неделю проходит через OpenRouter — рост в 5 раз за полгода

РАЗДЕЛ 03 · КАРТОЧКА МОДЕЛИ

DeepSeek V4 · DeepSeek (Ханчжоу)

Рабочая лошадка массового трафика: лучшая в мире цена за приемлемое фронтир-качество. V4 Flash — $0.14/$0.28 за миллион токенов при контексте 1M; попадание в кэш стоит $0.0028 — фактически бесплатно.

Оценка редакции (из 10; методика — стр. 02)



Сильные стороны

Ценовое лидерство без катастрофы в качестве: V4 Pro Max держит ~80.6% SWE-bench Verified — уровень, которого хватает для большинства продуктовых задач. Скорость ~60 токенов/сек — вдвое-втрое быстрее Kimi, что критично для real-time интерфейсов. Кэширование префиксов агрессивное и дешёвое: на чат-ботах с длинным системным промптом реальный счёт падает ещё в разы. MIT-лицензия на открытые веса.

Слабые стороны

По чистым агентным бенчмаркам уступает Kimi и GLM последнего поколения. Репутационный шлейф самый тяжёлый из всей пятёрки: именно приложение DeepSeek запрещали госорганы — для продаж в западный энтерпрайз имя в стеке придётся объяснять (как это делать — раздел 11). Ценовая политика волатильна: 31 мая закончилась 75%-я скидка на V4 Pro, и цена выросла в 4 раза за одну ночь.

Куда брать

Чат-боты поддержки, суммаризация, классификация, RAG-ответы, черновой кодинг — всё, где объём важнее последних процентов качества. Идеальная «дешёвая нога» в роутинге.

ВЕРДИКТ

Если продукт живёт на больших объёмах токенов — начинайте отсюда. Считайте экономику по цене после скидок, а не по промо.

РАЗДЕЛ 04 · КАРТОЧКА МОДЕЛИ

Kimi K2.6 · Moonshot AI (Пекин)

Чемпион долгих агентных задач: триллион параметров (MoE), 262K контекста, режим Agent Swarm с оркестрацией до 300 суб-агентов. Эталонный публичный прогон — 12+ часов автономной работы и 4000+ вызовов инструментов в одной сессии.

Оценка редакции (из 10; методика — стр. 02)

Сильные стороны

SWE-Bench Pro 58.6 — наравне с GPT-5.5 при цене ниже примерно на 80%; лидер LiveBench по коду (78.57) и агентному кодингу среди открытых моделей. Модель пост-тренирована именно под долгие многошаговые сессии с восстановлением после ошибок — то, на чём сыплются почти все конкуренты. Открытые веса (Modified MIT), официальные рецепты деплоя под vLLM, SGLang и KTransformers. На этой архитектуре Cursor построил Composer.

Слабые стороны

Бюджетной линейки нет — всё премиум: на массовом трафике экономика не сходится. Скорость 20–30 токенов/сек: для интерактивных интерфейсов это ощутимо медленно. Контекст 262K против 1M у DeepSeek и Qwen. Без кэширования префиксов ценовое преимущество над западными моделями испаряется — включать обязательно.

Куда брать

Автономные кодинг-агенты, мультиагентные пайплайны, задачи на часы работы без человека. Это «дорогая нога» роутинга — на неё уходят 10–20% самых сложных запросов.

ВЕРДИКТ

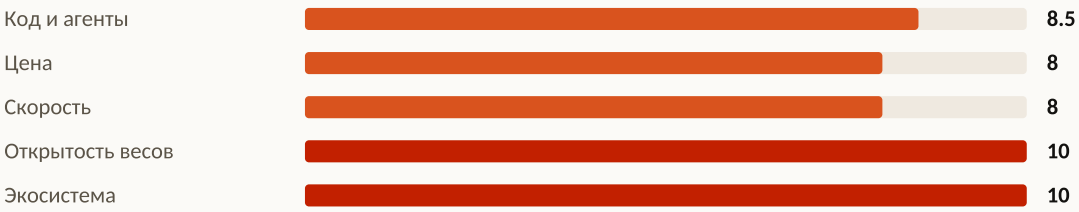
Лучшая открытая модель для серьёзной агентной работы. Брать точечно, на массовый трафик не ставить.

РАЗДЕЛ 05 · КАРТОЧКА МОДЕЛИ

Qwen 3.6 / 3.5 · Alibaba Cloud (Ханчжоу)

Самый полный стек из всех: линейка открытых моделей от 27B до 397B под Apache 2.0, эмбеддинги, мультимодальность, контекст 1M токенов и лучший на рынке tool-calling (48.2% MCPMark — первое место).

Оценка редакции (из 10; методика — стр. 02)



Сильные стороны

Единственный вендор, закрывающий весь диапазон: Qwen3.6-27B для локальной разработки на одной карте 24GB, Qwen3.5-122B-A10B для одного H100, флагман 397B для кластера — плюс облачный API ($0.50/$3.00 за Plus). Лучшая мультиязычность на рынке, включая уверенный русский. Apache 2.0 — никаких юридических сюрпризов. Самая надёжная работа с инструментами по протоколу MCP — фундамент для агентных продуктов. Доступен в AWS Bedrock как managed-вариант.

Слабые стороны

В чистом агентном кодинге уступает Kimi и GLM-5.1 верхнего эшелона. Флагман 3.6 Max-Preview — закрытый API без локального деплоя; открытая линейка отстаёт от него на шаг. Документация местами догоняет темп релизов.

Куда брать

Мультиязычные продукты, RAG-системы (эмбеддинги из той же экосистемы), MCP-агенты, и любой сценарий, где сегодня нужен API, а завтра — переезд на свои GPU без смены семейства моделей.

ВЕРДИКТ

Выбор по умолчанию, если важна гибкость: единственное семейство, с которым можно вырасти от ноутбука до кластера.

РАЗДЕЛ 06 · ДВЕ КАРТОЧКИ

GLM-5.1 и MiniMax M3

GLM-5.1

Zhipu AI / Z.ai (Пекин) · MIT-лицензия

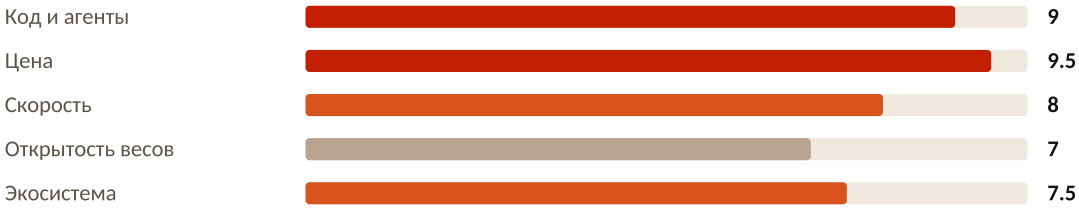
Сильное: лидер SWE-Bench Pro среди открытых моделей (58.4 — выше Claude Opus 4.6) и автономные задачи длительностью до 8 часов. Лицензия MIT — максимум свободы: инспектируй, модифицируй, продавай. GLM-4.7-Flash — полностью бесплатный API: лучший вариант «попробовать стек за ноль». На GLM построена модель Windsurf SWE-1.5; Vercel и Cerebras добавили её в свои платформы. **Слабое:** английский и русский слабее, чем у Qwen и DeepSeek (модель заточена под китайский: 94% на китайском сентимент-бенчмарке против 89% у GPT-4o); self-hosting тяжёлый — 595GB даже в INT4-квантовании, минимум 8×H100.

ВЕРДИКТ

Брать за MIT-свободу и фронтир-агентность; бесплатный Flash — лучшая песочница для старта.

MiniMax M3

MiniMax (Шанхай) · релиз июнь 2026



Сильное: самая дешёвая модель из всех, кто выбил больше 80% на SWE-bench Verified (80.5% при $0.30/$1.20 за миллион) — формально лучшее соотношение цена/качество в кодинге на рынке, в 17–20 раз дешевле западных аналогов того же уровня. Контекст 1M. Предшественник M2.5 неделями держал первое место по потреблению токенов на OpenRouter — модель проверена чужим продакшеном на триллионах токенов. **Слабое:** флагманские веса закрыты (открыты только младшие версии); экосистема и документация скромнее, чем у Alibaba; компания моложе — вендорский риск чуть выше.

ВЕРДИКТ

Лучший баланс цена/качество для кодинг-агентов, если не нужен self-hosting флагмана.

РАЗДЕЛ 07

Сводная таблица

Цены — официальные API за 1 млн токенов (вход/выход), июнь 2026. Для открытых моделей в скобках — цена у независимого хостера Together (ориентир, если прямой API не подходит по политикам).

МОДЕЛЬ	ЦЕНА IN/OUT, $/1M	КОНТЕКСТ	КЛЮЧЕВАЯ МЕТРИКА	ЛИЦЕНЗИЯ ВЕСОВ
DeepSeek V4 Flash	0.14 / 0.28 кэш-хит: 0.0028	1M	~60 ток/с; лучшая цена качественного инференса	MIT
DeepSeek V4 Pro	(2.10 / 4.40) скидка 75% истекла 31.05	1M	SWE-V ~80.6% (Max)	MIT
Kimi K2.6	~0.95+ / — (1.20 / 4.50)	262K	SWE-Pro 58.6 ≈ GPT-5.5; LiveBench Coding 78.57	Modified MIT
Qwen 3.6 Plus	0.50 / 3.00	1M	MCPMark 48.2% — лучший tool-calling	Apache 2.0 (откр. линейка)
GLM-5.1	(1.40 / 4.40) GLM-4.7-Flash: бесплатно	200K+	SWE-Pro 58.4 — лидер открытых	MIT
MiniMax M3	0.30 / 1.20	1M	SWE-V 80.5% — дешевейший в клубе 80%+	частично открытая
Справочно: GPT-класс	~2.50 / 10.00	до 1M	—	закрытая
Справочно: Claude-класс	до 25.00 за выход	1M	SWE-V 95% (Fable 5) — потолок рынка	закрытая

Как читать таблицу

Западные флагманы по-прежнему держат потолок качества (95% против ~80% на SWE-bench Verified) — если задача упирается в последние 15% точности, они незаменимы. Но большинство продуктовых задач живёт ниже этого потолка, и там решает цена за «достаточно хорошо». Разница между $0.28 и $10–25 за миллион выходных токенов — это разница между прибыльной и убыточной юнит-экономикой. Конкретные расчёты — на следующей странице.

РАЗДЕЛ 08

Юнит-экономика: три сценария

Цены за токен ничего не значат, пока не наложены на профиль нагрузки. Три типовых сценария, месяц работы, без учёта кэш-скидок (кэширование снижает все цифры — китайские сильнее прочих).

Сценарий A: чат-бот поддержки

100 тыс. диалогов/мес · ~3K токенов вход + 500 выход на диалог → 300M in / 50M out

DeepSeek V4 Flash: 300×$0.14 + 50×$0.28 = **$56/мес**; с кэшем системного промпта — реально $25–35. **GPT-класс** ($2.5/$10): $750 + $500 = **$1 250/мес**. Разница — **22–40 раз**. На этих объёмах западная цена — строка в бюджете, китайская — ошибка округления.

Сценарий B: RAG-ассистент по базе знаний

20 тыс. запросов/мес · 12K токенов контекста + 700 выход → 240M in / 14M out

DeepSeek V4 Flash: ≈ **$37/мес** (а с кэш-хитами на повторяющихся чанках документации — около $12–18). **Qwen 3.6 Plus:** $162/мес — если нужен лучший tool-calling и мультиязычность. **Западный фронтир** (~$3/$15): ≈ **$930/мес**. RAG — самый «кэшируемый» паттерн: выигрыш китайского стека здесь максимален.

Сценарий C: автономный кодинг-агент

агентные сессии — самый прожорливый паттерн: сотни тысяч токенов на задачу

Эталонный расчёт с одинаковым месячным workload: **Claude Fable 5 — $22 500/мес**, **DeepSeek V4 Flash — $252/мес**. Это два порядка. Практичная середина: **MiniMax M3** (~$1 000–1 500 на том же профиле при 80.5% SWE-V) или **Kimi K2.6** на критические задачи. Именно поэтому агентные нагрузки массово мигрировали на китайские модели — OpenRouter фиксирует, что больше половины всех выходных токенов платформы теперь генерируют агенты.

> Правильная архитектура — роутинг: 80–90% запросов на дешёвую модель, сложные 10–20% — на сильную. Это даёт качество флагмана при цене бюджетника.

22–40x
экономия на массовом чат-трафике

~$15
реальный месячный счёт RAG-системы с кэшем

89x
разница на агентном кодинге ($252 против $22 500)

30 юзкейсов: что и на чём строить

Каталог собран по принципу «задача → модель → почему именно она». Рекомендация — первый кандидат для прототипа; финальный выбор делает ваш eval-набор (раздел 14). Цены и маршруты подключения — в разделах 07–10.

Клиентский сервис и продажи

1	**Чат-бот поддержки первой линии**	**DeepSeek V4 Flash**	массовый трафик; с кэшем системного промпта — доли цента за диалог
2	**Голосовой бот колл-центра (распознавание → LLM → синтез)**	**DeepSeek V4 Flash**	60 ток/с — единственный из пятёрки, кто комфортно держит голосовую латентность
3	**Авторазбор и маршрутизация входящих тикетов**	**GLM-4.7-Flash**	классификация — простая задача; бесплатного API хватает на тысячи тикетов в день
4	**Персональные коммерческие предложения по данным CRM**	**Qwen 3.6 Plus**	лучший tool-calling: модель сама тянет поля сделки по MCP и собирает документ
5	**Анализ звонков отдела продаж: саммари, скоринг, возражения**	**DeepSeek V4 Flash**	ночные батчи транскриптов; кэш скриптов продаж режет цену ещё втрое

Контент и маркетинг

6	**SEO-статьи и лендинги на потоке (сотни в месяц)**	**DeepSeek V4 Flash**	цена решает: 200 статей по 2K токенов — меньше доллара
7	**Локализация продукта на 10+ языков**	**Qwen 3.6**	сильнейшая мультиязычность пятёрки, включая CJK и русский
8	**Описания тысяч товарных карточек для маркетплейса**	**MiniMax M3**	структурный вывод по шаблону при цене $1.20/M на выходе
9	**Персонализированные email-цепочки по сегментам**	**DeepSeek V4 Flash**	шаблон в кэш, переменные в запрос — рассылка на 100K адресов за центы
10	**Посты для соцсетей с tone-of-voice брендбука**	**Qwen 3.6 Plus**	брендбук целиком в контекст (1M токенов) вместо тонкой настройки

Разработка ПО

11	**Автономный агент на бэклоге: рефакторинг, тесты, мелкие баги**	**Kimi K2.6**	пост-тренирован под многочасовые сессии; эталон — 12 часов и 4000+ вызовов инструментов
12	**Code review каждого pull request в CI**	**MiniMax M3**	80.5% SWE-bench Verified по цене, при которой ревью каждого PR ничего не стоит
13	**Генерация и поддержка автотестов**	**MiniMax M3**	однотипная объёмная работа — профиль модели идеален
14	**Миграция легаси-кода (язык → язык, фреймворк → фреймворк)**	**Kimi K2.6**	публичный эталонный кейс Moonshot — именно портирование кодовой базы
15	**Внутренний copilot, код не покидает контур**	**Qwen3.6-27B self-host**	работает на одной карте 24GB; Apache 2.0 — без вопросов от юристов

Данные, операции, вертикали

Данные и документы

16	**RAG-ассистент по базе знаний компании**	**DeepSeek V4 Flash**	повторяющиеся чанки документации почти целиком уходят в кэш по $0.0028/M
17	**Извлечение полей из счетов, накладных, договоров**	**Qwen 3.6 Plus**	дисциплина структурного вывода (JSON) — фирменная сильная сторона семейства
18	**Суммаризация документов на сотни страниц**	**Qwen 3.6 Plus**	контекст 1M: годовой отчёт помещается целиком, без нарезки и потери связей
19	**Юридический ассистент: анализ договоров на риски**	**GLM-5.1 self-host**	конфиденциальность требует своего контура; MIT-лицензия упрощает согласование
20	**Разбор и тегирование документального архива**	**GLM-4.7-Flash**	миллионы строк классификации по нулевому тарифу — грех не использовать

Операции и внутренние инструменты

21	**Ассистент по регламентам и базе знаний для сотрудников**	**DeepSeek V4 Flash**	внутренние нечувствительные данные — прямой API допустим, цена символическая
22	**Еженедельные сводки из метрик и дашбордов**	**Qwen 3.6 Plus**	tool-calling к BI-системе по MCP: модель сама собирает цифры и пишет отчёт
23	**HR: скрининг резюме и черновики вакансий**	**Qwen3.6-27B self-host**	персональные данные кандидатов — повод держать инференс у себя
24	**Переписка с китайскими поставщиками: перевод + саммари**	**GLM-5.1**	китайский — родной язык модели (94% на сентимент-бенчмарке против 89% у GPT-4o)
25	**Аналитик по запросу: вопрос на русском → SQL → ответ**	**Qwen 3.6 Plus**	лучшая надёжность цепочки «вопрос → инструмент → проверка → ответ»

Продукты и вертикали

26	**EdTech: тьютор с пошаговой проверкой решений**	**DeepSeek V4**	сильная математика в reasoning-режиме при массовой цене
27	**E-commerce: рекомендации и Q&A на карточке товара**	**MiniMax M3**	миллионы микрозапросов; баланс качества и цены — точно его профиль
28	**Медиа: мониторинг китайских источников и дайджесты**	**Qwen 3.6 + GLM**	родной китайский + длинный контекст; на этой связке работает и наш канал
29	**Финтех: объяснения скоринга и антифрод-комментарии**	**GLM-5.1 self-host**	регулятор требует аудит-трейл и контроль весов — только свой контур
30	**Игры: живые диалоги NPC в реальном времени**	**DeepSeek V4 Flash**	латентность и цена — два требования жанра, оба закрыты лучше всех

> Закономерность каталога: DeepSeek берёт объём, Qwen — структуру и языки, Kimi — длинные агентные задачи, MiniMax — баланс, GLM — китайский язык, свободу лицензии и бесплатный старт.

РАЗДЕЛ 10

Четыре маршрута подключения

МАРШРУТ	КАК ЭТО РАБОТАЕТ	КОГДА ВЫБИРАТЬ
1. Прямой API	Все пять вендоров дают OpenAI-совместимые endpoint'ы: меняете baseURL и ключ в существующем SDK — код не трогаете. Минуты на интеграцию	Прототипы, внутренние инструменты, рынки без чувствительности к данным в КНР
2. Агрегаторы	OpenRouter, Vercel AI Gateway, Helicone, Portkey: один ключ — 400+ моделей, единый биллинг, фолбэки и роутинг из коробки	Роутинг-стратегии, быстрые A/B моделей, защита от отказа одного вендора
3. Западные облака	AWS Bedrock (эндпоинт bedrock-mantle: DeepSeek, Kimi, Qwen, GLM, MiniMax — ~24 открытые модели), Azure, Together, Fireworks: веса китайские, серверы и договор — западные	Энтерпрайз-комплаенс: данные не покидают выбранную юрисдикцию, SLA и поддержка от облака
4. Self-hosting	Открытые веса + vLLM/SGLang на своих или арендованных GPU. Полный контроль, фиксированная цена за железо вместо цены за токен	Регулируемые отрасли, большие стабильные объёмы, авиагэп-среды. Детали — стр. 12

Специфика подключения из России

Парадокс момента: китайские API — единственный фронтир-AI, доступный из РФ напрямую, без VPN и без риска внезапной блокировки аккаунта по географии. Кабинеты DeepSeek, Moonshot, Zhipu и Alibaba Cloud работают с российскими IP штатно. Оплата — основное трение: прямые кабинеты принимают международные карты (решается картой зарубежного банка или корпоративным аккаунтом через юрлицо в дружественной юрисдикции); часть агрегаторов принимает криптовалюту. Для компаний практичнее всего связка: договор с Alibaba Cloud или китайским реселлером + оплата в юанях по обычному ВЭД-контракту — это штатная внешнеторговая операция.

Мини-рецепт миграции

проверено сообществом, типовой путь

1) Собираете 50–100 реальных запросов вашего продукта в eval-набор. 2) Через OpenRouter прогоняете их по 3–4 китайским моделям и текущей западной. 3) Сравниваете качество вслепую + считаете цену. 4) Переключаете baseURL для сегмента трафика 10%, неделю смотрите метрики. 5) Раскатываете роутинг. Типовой срок всего цикла — одна рабочая неделя, инженерных изменений — почти ноль.

РАЗДЕЛ 11

Self-hosting: железо под модель

Главное правило: выбор модели для self-hosting — это на 90% выбор по имеющемуся железу. Карта соответствий на июнь 2026:

ЖЕЛЕЗО	ЧТО ВЛЕЗАЕТ	КОММЕНТАРИЙ
1× GPU 24GB RTX 4090/5090	Qwen3.6-27B	Локальная разработка, прототипы, приватные ассистенты. Удивительно пристойный код
1× H100 (80GB)	Qwen3.5-122B-A10B; DeepSeek V4-Flash (квант.)	MoE-архитектура: 122B параметров при 10B активных — продакшн отдела на одной карте
4× H100 (320GB)	GLM-4.7	Входной билет в кластерный класс; сильный кодинг
8× H100/H200 ~640GB+	Kimi K2.6 INT4 (594GB); GLM-5.1 INT4 (595GB)	Фронтир у себя. 8×H200 — комфорт, 8×H100 — впритык; альтернатива — 4×B200 (AWQ ~630GB)
16× H100 (2 узла)	Kimi K2.6 / GLM-5.1 в FP16 (~1.5–2TB)	Полная точность — нужна редко: INT4-кванты этих моделей почти без потерь
Воркстейшн 1.5TB RAM + 1 GPU	K2.6 GGUF ~350GB (UD-Q2_K_XL)	~10 ток/с через KTransformers: годится для оффлайн-задач и разработки, не для трафика

Софтверный слой

vLLM — дефолт для высоконагруженного OpenAI-совместимого сервинга. **SGLang** — быстрее на агентных нагрузках с повторяющимися префиксами (механизм RadixAttention); если строите агентов — начинайте с него. **KTransformers** — гибрид CPU+GPU для «бедного» железа. Для агентных нагрузок включение prefix-кэширования обязательно: без него теряется главный экономический смысл. Не хотите операционки вообще — управляемый средний путь: AWS Bedrock (Project Mantle) или Together с выделенным деплоем.

Когда self-hosting окупается

Грубая граница: аренда 8×H200 стоит ~$15–25 тыс./мес. Если ваш API-счёт на премиальных моделях стабильно выше — считайте переезд; если основной объём идёт на дешёвый Flash-класс по $0.14 — почти наверняка не окупится никогда. Self-hosting выбирают не ради экономии, а ради контроля: данные, версии весов, независимость от ценовых сюрпризов вендора.

РАЗДЕЛ 12

Юридика и риски: матрица решений

Самая частая ошибка в спорах о китайских моделях — смешивать **приложение** и **веса**. Запреты правительств (Австралия, Чехия, Италия, ведомства США) касаются приложения DeepSeek и его серверов: по privacy policy данные пользователей хранятся в КНР, где закон обязывает компании сотрудничать со спецслужбами. К открытым весам, запущенным на вашем железе или в западном облаке, эта претензия не применима — данные туда физически не уходят.

> Microsoft запретила сотрудникам приложение DeepSeek — и одновременно хостит модель DeepSeek в Azure. Это не лицемерие, а точное понимание, где именно живёт риск.

Четыре категории риска и их закрытие

Данные (residency)	Прямой API = данные на серверах в КНР: непригодно для персональных данных ЕС (GDPR) и чувствительных данных США. Закрывается маршрутами 3–4: Bedrock/Together/self-host
Цензура в весах	Контент-фильтры по политическим темам (Тайвань, Синьцзян) вшиты в обучение и сохраняются при self-hosting. Для типового B2B-продукта нерелевантно, но прогоните свой домен на тесте — особенно медиа и образование
Регуляторика	Письма Конгресса Cursor/Airbnb (апрель 2026) — сигнал направления. Если рынок — госсектор/оборона/критическая инфраструктура США, закладывайте сценарий принудительной миграции: абстрагируйте модельный слой
Операционный	Ценовая волатильность (промо DeepSeek ×4 за ночь), жёсткие rate limits бесплатных тиров, версии без предупреждения. Закрывается фиксацией версий, фолбэками через агрегатор и запасом в юнит-экономике

Матрица: что можно вашему продукту

ПРОФИЛЬ ПРОДУКТА	ДОПУСТИМЫЙ МАРШРУТ
Внутренние инструменты, нечувствительные данные	**Любой, включая прямой API**
B2C/B2B с персданными ЕС или США	**Только западный хостинг весов или self-host**
Финансы, медицина, госконтракты	**Self-host + аудит-трейл + фиксация весов**
Российский рынок	**Любой; прямой API — без юридических ограничений**

РАЗДЕЛ 13

Семь антипаттернов

Каждый пункт — реальные грабли, на которые уже наступили команды до вас. Чтение этой страницы дешевле любого из этих уроков.

1 **Персданные через прямой API**

Гнать данные клиентов из ЕС/США напрямую в китайский endpoint — кратчайший путь к GDPR-штрафу и потере энтерпрайз-сделок. Правильно: западный хостинг весов или self-host (раздел 11).

2 **Юнит-экономика на промо-ценах**

DeepSeek поднял цену V4 Pro в 4 раза за одну ночь по окончании скидки. Считайте бизнес-кейс по полной цене, промо — приятный бонус, а не основа модели.

3 **Kimi на массовый трафик**

Премиальная цена без бюджетной линейки и 20–30 ток/с убивают экономику и UX чат-ботов. Kimi — скальпель для сложных агентных задач, а не лопата для объёма.

4 **Игнор кэширования префиксов**

В агентных нагрузках системный промпт и история пересылаются каждым вызовом. Без prefix-кэша вы платите за одно и то же десятки раз — и теряете весь ценовой выигрыш стека.

5 **Одна модель на всё**

Моногамия с любой моделью — переплата. Роутинг «дешёвая на массу + сильная на сложное» даёт качество флагмана по цене бюджетника; агрегаторы делают это из коробки.

6 **Вера вендорским бенчмаркам**

Цифры в релизах — маркетинг по чужим задачам. 50–100 реальных запросов вашего продукта как eval-набор скажут больше, чем любой лидерборд, и соберутся за день.

7 **Self-host без фиксации версии**

Веса обновляются, поведение меняется молча. Фиксируйте хеш весов, держите регрессионный eval и план на случай дефицита GPU — иначе «контроль» self-hosting'а иллюзорен.

РАЗДЕЛ 14

Чеклист внедрения

- [ ] Собрали eval-набор из 50–100 реальных запросов продукта
- [ ] Прогнали 3–4 кандидата + текущую модель вслепую, сравнили качество и цену
- [ ] Посчитали юнит-экономику по полным (не промо) ценам с учётом кэширования
- [ ] Классифицировали данные: что можно в прямой API, что только в западный хостинг/self-host
- [ ] Выбрали маршрут подключения (прямой / агрегатор / облако / self-host) под матрицу рисков
- [ ] Включили prefix-кэширование и проверили его срабатывание в счетах
- [ ] Настроили роутинг и фолбэк на вторую модель через агрегатор
- [ ] Зафиксировали версии моделей/весов; завели регрессионный eval на обновления
- [ ] Прогнали чувствительные для домена темы (цензурные фильтры) на тестовом наборе
- [ ] Раскатили на 10% трафика, неделю сверяли метрики, затем — полный переезд

ГЛОССАРИЙ

MoE	Mixture of Experts — архитектура, активирующая лишь часть параметров на запрос: триллионная модель по цене инференса средней
SWE-bench Verified / Pro	бенчмарки решения реальных задач из GitHub-репозиториев; Pro — усложнённая закрытая версия от Scale AI
MCP	Model Context Protocol — открытый протокол подключения моделей к инструментам и данным; MCPMark — бенчмарк надёжности такой работы
Prefix-кэширование	повторное использование уже обработанной части промпта (системные инструкции, история); кэш-хиты у DeepSeek в 50 раз дешевле обычного входа
INT4 / AWQ / GGUF	форматы квантования весов: сжатие модели в 2–4 раза с малой потерей качества — то, что делает self-hosting фронтира реальным
vLLM / SGLang / KTransformers	открытые движки инференса: универсальный, агентно-оптимизированный и гибридный CPU+GPU соответственно
OpenRouter	агрегатор 400+ моделей с единым API и биллингом; его статистика — лучший публичный индикатор реального продакшн-использования
AWS Bedrock (Project Mantle)	управляемый сервинг открытых моделей в облаке Amazon: китайские веса на западной инфраструктуре
Роутинг	автоматическое распределение запросов между моделями по сложности/цене — базовый паттерн экономики 2026 года
Eval-набор	ваш собственный тестовый набор реальных запросов с эталонными ответами — единственный бенчмарк, которому стоит верить

Источники: официальные прайс-листы DeepSeek, Moonshot, Zhipu/Z.ai, Alibaba Cloud, MiniMax; Together.ai; OpenRouter (статистика использования); Bloomberg, TechCrunch, Reuters, SCMP, kr-asia, 36Kr; LiveBench, SWE-bench, MCPMark, Artificial Analysis; гайды деплоя vLLM/SGLang и официальные model cards Hugging Face. Данные актуальны на 11 июня 2026 г. Цены меняются часто — перепроверяйте перед бюджетированием.

技

1,4 млрд человек строят будущее. Мы переводим.

Технологические новости из Китая — с разбором,
контекстом и без воды. Каждый день.

@news_chinatech

Отчёт распространяется свободно при сохранении ссылки на канал. Предыдущие отчёты серии — в закрепе канала.