Стоимость API языковых моделей: США · Китай · Россия

Цены, бенчмарки, LMArena и сводный параметр «интеллект на доллар»

Дата сборки: 26 июня 2026

Аналитик: ResearchGate · skill deep-research (deep mode)

Цель отчёта: Сравнить стоимость API текстовых LLM трёх юрисдикций с качеством по Artificial Analysis Intelligence Index + LMArena Elo, включая подписки и объём токенов

Режим: deep · фокус — работа с текстом · цены в USD за 1M токенов · курсы $1 ≈ 80 ₽, $1 ≈ 7,2 CNY

Краткое резюме

К середине 2026 года рынок текстовых LLM окончательно распался на **три почти не пересекающихся контура**, и главный вывод исследования именно про их разделение труда, а не про «кто лучше вообще».

1. **США держат абсолютный потолок качества, но дорого.** Composite-лидеры (живой снимок лидербордов на 26.06.2026) — **Claude Fable 5** (сводный балл 98/100, #1 и в AA Index, и в LMArena), **Claude Opus 4.8** (89), **GPT-5.5** (88), **GPT-5.4** (83), **Gemini 3.5 Flash** (82). За потолок платят премию: blended-цена топ-флагманов США — $4,5–20 за 1M токенов, в 8–40 раз дороже сопоставимых по качеству китайских моделей.
2. **Китай выигрывает «интеллект на доллар» с разгромным счётом.** В топ-10 по сводному параметру **Value** (качество ÷ цена) — 8 из 10 строк китайские. **DeepSeek-V4-Pro** даёт 72,5 балла качества за blended $0,54 (value 133) — уровень почти-флагмана за треть цены Sonnet. **GLM-5.2** (Zhipu) — 82 балла качества за $2,15: ближайший преследователь топа США по абсолютному качеству и при этом в 5–9 раз дешевле. **Qwen3.7-Max** — 78 баллов за $0,52, **MiniMax M3** — 71 за $0,44, **DeepSeek-V4-Flash** — 65 баллов за $0,175 (рекордный value 369).
3. **Россия — отдельный локальный рынок без выхода на мировые лидерборды.** Ни YandexGPT, ни GigaChat не присутствуют в LMArena и Artificial Analysis; их качество измеряется русским бенчмарком **MERA**, где GigaChat 2 Max (0,67) по самооценке Сбера обходит GPT-4o. По цене RU-API ближе к среднему сегменту США/мира: YandexGPT 5.1 Pro ≈ $5/1M (единая ставка вход=выход), GigaChat 2 Max ≈ $8/1M. Уникальное преимущество — соответствие 152-ФЗ, локализация данных и рубли в кассе.
4. **Поколения сменились быстрее, чем успевают писать обзоры.** На июнь-2026 актуальны GPT-5.5/5.4, Claude Opus 4.8, Gemini 3.1 Pro / 3.5 Flash, Grok 4.3, DeepSeek V4, GLM-5.2, Qwen 3.7, Kimi K2.6 — а имена «GPT-5 / Gemini 2.5 / DeepSeek R1 / Claude 4.5», которые ещё цитирует большинство публикаций, отстали на 1–2 поколения.
5. **Подписки почти нигде не считаются в токенах.** ChatGPT, Claude, Gemini в 2026 перешли на «скользящее окно + недельный лимит» и **не публикуют токенные квоты** — все оценки объёма здесь расчётные. Единственный честно токен-номинированный потребительский тариф — **GigaChat Freemium: 1 000 000 токенов на 12 месяцев** (~83K/мес).

Practical bottom line. Нужен максимум интеллекта на сложном тексте и цена не критична → Opus 4.8 / GPT-5.5. Нужна та же лига качества при минимальной цене → **DeepSeek-V4-Pro или GLM-5.2**. Нужен дёшево-массовый текст (классификация, извлечение, черновики) → **DeepSeek-V4-Flash, Qwen-Flash, Gemini 2.5 Flash-Lite** ($0,06–0,18 blended). Нужны 152-ФЗ и данные в РФ → YandexGPT / GigaChat.

Введение: предмет, рамки и метод

Предмет. Стоимость доступа к большим языковым моделям через API для задач **работы с текстом** (генерация, суммаризация, извлечение, классификация, переписывание, диалог) в трёх юрисдикциях

— США, Китай, Россия — и сопоставление этой стоимости с качеством моделей через единый сводный параметр.

Что включено. Текстовые модели (text-in / text-out). Для каждой — цена входа и выхода за 1M токенов, контекстное окно, рассуждающий (reasoning) режим, кэш-скидки. Отдельно — потребительские подписки и включённый в них объём токенов. Мультимодальность (картинки/аудио/видео), эмбеддинги и файн-тюнинг — за рамками.

Как считаем «сводный параметр». Запрос требовал «оценку по бенчмаркам и по lmarena» одним числом. Мы строим его из двух независимых осей качества:

- **Artificial Analysis Intelligence Index** — композит из MMLU-Pro, GPQA Diamond, AIME, LiveCodeBench и др. Это лучшая единая ось *способности решать задачи*. Шкала пересмотрена в начале 2026 (текущий максимум ~56 у Opus 4.8).
- **LMArena Elo (text)** — рейтинг *человеческих предпочтений* в слепых сравнениях. Лучше ловит «приятность» ответа и chat-качество.

Формулы (полностью прозрачны, скрипт `compute_scores.py` в папке):

```
blended = (3·вход + 1·выход) / 4        # конвенция Artificial Analysis, 3:1 вход:выход
Q_aa    = AA_Index / 56 · 100           # нормировка к топу
Q_arena = (Elo − 1300) / 220 · 100      # обрезка 0..100
Quality = 0,6·Q_aa + 0,4·Q_arena        # если есть обе оси; иначе доступная
Value   = Quality / blended             # «интеллект на доллар»
```

Вес 0,6/0,4 в пользу Intelligence Index — потому что для «работы с текстом» способность важнее симпатии. Российские модели в основной композит **не входят** (нет данных AA/Arena) и разбираются отдельно по MERA.

Confidence-маркировка (по методологии ResearchGate). Каждая цифра помечена: **High** — официальная страница тарифов/документации; **Med** — агрегатор (Artificial Analysis, OpenRouter, pricepertoken) или вторичный обзор; **Low** — оценка, промо-ставка или конфликт источников. Цены LLM в 2026 меняются ежемесячно — отчёт фиксирует снимок на 26.06.2026.

Ключевые допущения (вынесены явно, как требует методология): - Курсы: $1 = 80 ₽, $1 = 7,2 CNY. - Leaderboard-данные — снимок ~23.06.2026; рейтинги переранжируются еженедельно, CI у топа перекрываются. - Подписочные токен-оценки — теоретический потолок (cap × число окон × ~1500 ток/сообщение); реальное потребление обычно 10–25% от него.

---

1. Общая картина: рынок сменил поколение

Главное, что нужно понять перед таблицами: **между публикацией большинства обзоров и сегодняшним днём сменилось 1–2 поколения моделей.** Сопоставление «старое имя → актуальная модель» на 26.06.2026:

Регион	Что цитируют обзоры	Что реально актуально (июнь-2026)
OpenAI	GPT-5, o3	**GPT-5.5 / GPT-5.4** (+ mini/nano), GPT-5.3-codex
Anthropic	Claude 3.5/4.1	**Claude Opus 4.8**, Sonnet 4.6, Haiku 4.5
Google	Gemini 2.5 Pro	**Gemini 3.1 Pro / 3.5 Flash** (2.5 — прошлое поколение)
xAI	Grok 4	**Grok 4.3**
DeepSeek	V3 / R1	**DeepSeek-V4** (Pro / Flash)
Zhipu	GLM-4.5	**GLM-5.2 / 5.1**, GLM-4.7
Alibaba	Qwen-Max / Qwen3	**Qwen3.7-Max**, Qwen-Plus/Flash
Moonshot	Kimi K2	**Kimi K2.6 / K2.7**
Baidu	ERNIE 4.5 / X1	**ERNIE 5.0** (+ 4.5-Turbo, X1-Turbo)

Инсайт. Скорость смены поколений сама по себе — фактор стоимости: цена «вчерашнего флагмана» падает в 2–3 раза в момент выхода нового (так Яндекс срезал YandexGPT втрое в августе-2025, Сбер — GigaChat втрое в феврале-2026, OpenAI увёл GPT-4.1/o3 с витрины). Поэтому стратегия «брать прошлое поколение» почти всегда даёт лучший value, чем гнаться за свежим флагманом.

---

2. США: цены на API (текст)

Доллары за 1M токенов. «Cached» — цена чтения из кэша. Источники официальных страниц = High.

OpenAI

На живой витрине осталось только поколение GPT-5.x; GPT-4.1 и o-серия отвечают по API, но убраны из публичной таблицы (legacy → Med).

Модель	Вход	Выход	Cached	Контекст	Примечания	Conf
gpt-5.5	$5,00	$30,00	$0,50	~1M	Флагман, ~апр-2026	High
gpt-5.5-pro	$30,00	$180,00	—	~1M	Макс-reasoning	High
gpt-5.4	$2,50	$15,00	$0,25	~1M	Рабочая лошадка, reasoning	High
gpt-5.4-mini	$0,75	$4,50	$0,075	~1M		High
gpt-5.4-nano	$0,20	$1,25	$0,02	~1M	Самая дешёвая у OpenAI	High
gpt-5.3-codex	$1,75	$14,00	$0,175	~1M	Под код	High
gpt-4.1 (legacy)	$2,00	$8,00	$0,50	1M	Не на витрине	Med
o4-mini (legacy)	$1,10	$4,40	$0,275	200K	Reasoning	Med

Anthropic (Claude) — всё High (официальная страница)

Cache-read = 0,1× входа. Окно 200K (1M для Sonnet 4.6 и Opus 4.6–4.8 по стандартной ставке).

Модель	Вход	Выход	Cached	Контекст	Примечания	Conf
Claude Fable 5	$10,00	$50,00	$1,00	1M	**#1 в LMArena (1508) и AA Index (60)** — топ рынка	High
Claude Opus 4.8	$5,00	$25,00	$0,50	1M	Флагман, ~28.05.2026; Fast-режим $10/$50	High
Claude Opus 4.7	$5,00	$25,00	$0,50	1M	Reasoning	High
Claude Sonnet 4.6	$3,00	$15,00	$0,30	1M	Reasoning	High
Claude Haiku 4.5	$1,00	$5,00	$0,10	200K	Быстрая/дешёвая	High
Claude Opus 4.1 (deprec.)	$15,00	$75,00	$1,50	200K	Старая цена флагмана	High

> ⚠️ *Opus 4.7+ используют новый токенизатор, потребляющий до ~35% больше токенов на тот же текст — реальная стоимость на практике выше номинала. Для моделей 4.6+ опция* `inference_geo=US` *добавляет множитель 1,1×.*

Google Gemini — High (официальная страница)

Часть моделей — тарифицируется ступенчато по размеру промпта (≤200K vs >200K входных токенов; ниже показана нижняя ступень / верхняя).

Модель	Вход	Выход	Cached	Контекст	Примечания	Conf
Gemini 3.1 Pro	$2,00 / $4,00	$12,00 / $18,00	$0,20	1M+	Ступенчато ≤200K/>200K; reasoning	High
Gemini 3.5 Flash	$1,50	$9,00	$0,15	1M	Новейший Flash	High
Gemini 3.1 Flash-Lite	$0,25	$1,50	$0,025	1M		High
Gemini 2.5 Pro	$1,25 / $2,50	$10,00 / $15,00	$0,125	1M+	Прошлое поколение	High
Gemini 2.5 Flash	$0,30	$2,50	$0,03	1M		High
Gemini 2.5 Flash-Lite	$0,10	$0,40	$0,01	1M	Дешевейшая у Google	High

xAI Grok / Mistral / прочие

Модель	Провайдер	Вход	Выход	Контекст	Conf
Grok 4.3	xAI	$1,25	$2,50	1M	High
Grok 4.1-fast	xAI	$0,20	$0,50	2M	Low (агрегатор)
Mistral Large 3	Mistral	$0,50	$1,50	128K+	High
Mistral Medium 3.5	Mistral	$1,50	$7,50	128K	High
Mistral Small 4	Mistral	$0,15	$0,60	128K	High
Command A	Cohere	$2,50	$10,00	256K	Low (нет на витрине)
Nova Pro	Amazon	$0,80	$3,20	300K	Med
Nova Lite	Amazon	$0,06	$0,24	300K	Med
Llama 4 Maverick	Meta (3rd-party)	$0,15	$0,60	1M	Low
Llama 4 Scout	Meta (3rd-party)	$0,08	$0,30	10M	Low

Инсайт по США. Внутри США разброс цен — двузначный: от $0,04 blended (Nova Lite) до $11+ (GPT-5.5). Самое выгодное соотношение цена/качество среди американских — **Grok 4.3** ($1,56 blended при качестве 78,5) и **Gemini 3.1 Pro** ($4,5 при 86). Флагманы Opus 4.8 и GPT-5.5 — это «купить последние 10–15 баллов качества» с премией ×3–5.

---

3. Китай: цены на API (текст)

Конвертация 7,2 CNY/$. Цены — стандартные (cache-miss) за 1M токенов. DeepSeek и Zhipu (z.ai) публикуют цены в USD на международных эндпойнтах = High; Alibaba/Baidu/ByteDance/Tencent — в CNY на материковых консолях (агрегатор/пресса = Med/Low).

DeepSeek — High (официально, в USD)

Модель	Вход (miss)	Вход (hit)	Выход	Контекст	Примечания	Conf
DeepSeek-V4-Flash	$0,14	$0,0028	$0,28	1M	Самый дешёвый frontier-класс; объединил chat+reasoner	High
DeepSeek-V4-Pro	$0,435	$0,0036	$0,87	1M	Флагман-reasoning (вероятно промо; лист до $1,74/$3,48)	High

> *Подпись DeepSeek: cache-hit ~в 50× дешевле входа — главная экономия на повторяющихся промптах/RAG.*

Zhipu / GLM — High (z.ai, в USD)

Модель	Вход	Выход	Контекст	Conf
GLM-5.2	$1,40	$4,40	~200K	High
GLM-5	$1,00	$3,20	~200K	High
GLM-4.7	$0,60	$2,20	~200K	High
GLM-4.7-FlashX	$0,07	$0,40	128K	High
GLM-4.5-Air	$0,20	$1,10	128K	High
GLM-4.7-Flash	Free	Free	128K	High

Alibaba Qwen / Moonshot Kimi / Baidu / ByteDance / прочие — Med/Low

Модель	Провайдер	Вход	Выход	Контекст	Conf
Qwen3.7-Max	Alibaba	$0,35	$1,04	256K	Med
Qwen-Plus	Alibaba	$0,40	$1,20	128K–1M	Med
Qwen-Flash	Alibaba	$0,033	$0,13	до 1M	Low
Kimi K2.6	Moonshot	$0,95	$4,00	256K	Med
Kimi K2.5	Moonshot	$0,60	$3,00	256K	Med
ERNIE 4.5	Baidu	$0,55	$2,22	128K	Med
ERNIE X1 (reasoning)	Baidu	$0,28	$1,10	32K+	Med
ERNIE 4.5-Turbo	Baidu	$0,11	$0,44	128K	Med
Doubao-Seed-1.6	ByteDance	$0,11	$1,11	256K	Med
MiniMax M2	MiniMax	$0,255	$1,00	197K	Med
MiniMax M1	MiniMax	$0,40	$2,20	1M	Med
Hunyuan-TurboS	Tencent	$0,11	$0,28	256K	Med
Hunyuan-T1 (reasoning)	Tencent	$0,14	$0,56	256K	Low-Med
Yi-Lightning	01.AI	$0,14	$0,14	16K	Low

Инсайт по Китаю. Китайский рынок устроен принципиально иначе: (1) **агрессивные кэш-скидки** (DeepSeek –98% на cache-hit, Qwen –90%, Kimi –83%); (2) **ступенчатая тарификация по длине контекста** — цены выше указанных при >128–256K; (3) **batch –50%**. Реальная стоимость в продакшене с кэшем и батчем у DeepSeek/Qwen падает ещё в 2–3 раза относительно номинала. Именно поэтому весь топ Value — китайский.

---

4. Россия: цены на API (текст)

Особенность RU-рынка: **единая цена за 1000 токенов (вход = выход)**, асинхронный режим = половина синхронного. Курс 80 ₽/$.

Yandex — YandexGPT (Yandex Cloud AI Studio)

Модель	₽/1000 ток (синх)	Вход $/1M	Выход $/1M	Контекст	Примечания	Conf
YandexGPT 5.1 Pro	0,40 (асинх 0,20)	$5,00	$5,00	32K	Флагман; цену срезали ×3 при релизе 28.08.2025	High*
YandexGPT 5 Lite	0,20 (асинх 0,10)	$2,50	$2,50	8–32K	Лёгкая	Med

* Официальная страница тарифов Яндекса на 26.06.2026 за CAPTCHA; цифры — из пресс-релиза Яндекса о снижении цены втрое + RU-обзоры. По «предыдущему 5 Pro» источники расходятся (1,20 vs 0,80 ₽).

Сбер — GigaChat API (тарифы для юрлиц, с 01.02.2026) — High

Модель	₽/1000 ток (синх)	Вход $/1M	Выход $/1M	Контекст	Примечания	Conf
GigaChat 2 Lite	0,065	$0,81	$0,81	128K	Дешевейший платный RU-API; пакет 1 млрд ток = 65 000 ₽	High
GigaChat 2 Pro	0,50	$6,25	$6,25	128K		High
GigaChat 2 Max	0,65	$8,13	$8,13	128K	Флагман	High

> *Сбер в феврале-2026 снизил цены ~втрое. Free tier: 1 000 000 токенов на 12 мес. Pay-as-you-go: мин. платёж 600 ₽/мес (если был расход). Доступны пакеты токенов и асинхронный режим (–50%).*

T-Bank / MTS / прочие

Модель	Провайдер	Статус	Conf
T-Pro 2.0 (32B), T-Lite	T-Bank	**Open-weight, Apache 2.0, бесплатно** на HuggingFace; своего платного API нет — self-host/агрегатор	High
Cotype (Pro)	MTS AI	~1,1 ₽/1000 ток (~$13,75/1M), единая; публичного self-serve прайса нет, доступ по запросу	Low
VK и др.	—	Публичного тарифицируемого LLM-API в открытом доступе на 06.2026 не найдено	Med

Инсайт по России. RU-цены не «дешевле мира» — YandexGPT 5.1 Pro ($5/1M) и GigaChat 2 Max ($8/1M) сопоставимы со средним сегментом США (Sonnet $3–15, Gemini Pro), но при этом *по абсолютному качеству ближе к китайскому/американскому среднему классу, а не к флагманам*. Экономический смысл RU-моделей — не цена и не максимум интеллекта, а **152-ФЗ, локализация ПД, оплата в**

рублях без валютных и санкционных рисков, интеграция в экосистемы (Yandex Cloud, SberDevices). Для задач, где это критично, альтернатив фактически нет. Лучший value среди RU — GigaChat 2 Lite ($0,81/1M).

---

5. Качество: бенчмарки, LMArena, Intelligence Index

Живой снимок DOM лидербордов на 26.06.2026 (artificialanalysis.ai + lmarena.ai, прочитано Playwright). AA Index — на пересмотренной шкале (топ = 60 у Fable 5). Отдельные бенчмарки (GPQA/AIME/coding) — launch-era значения. Где AA даёт варианты по «усилию рассуждения» (low/med/high/xhigh/max) — показан старший стандартный.

Модель	Рег	AA Index	LMArena Elo (ранг)	GPQA Diamond	AIME 2025	Coding (SWE-bench V.)	Conf
Claude Fable 5	US	**60**	**1508 (#1)**	—	high	high	High
Claude Opus 4.8	US	56	1484 (#9 think)	93,6%	high	88,6%	High
GPT-5.5	US	55	1481 (#10 high)	~88%	~95%	~75%	High
GPT-5.4	US	51	1478 (#12 high)	85–89%	94,6%	74,9%	High
GLM-5.2	CN	51	1470 (#25)	~83%	91%	64,2% (GLM-4.5)	High
Gemini 3.5 Flash	US	50	1476 (#13)	—	—	—	High
Sonnet 4.6	US	47 (max)	1472 (#23)	83,4% (4.5)	~70%	77,2%	High
Gemini 3.1 Pro	US	46	1486 (#7)	—	—	—	High
Qwen3.7-Max	CN	46	1475 (#16)	81,1%	92,3%	~70% (LiveCodeBench)	High
DeepSeek-V4-Pro	CN	44	1457 (#39)	79,9% (V3.2)	89,3% (V3.2)	—	High
MiniMax M3	CN	44	1447 (#51)	~70% (self)	~86% (self)	—	High
Kimi K2.6	CN	43	1461 (#33)	75,1%	—	65,8%	High
DeepSeek-V4-Flash	CN	40	1435 (#66)	—	—	—	High
GPT-5.4-mini	US	40	1449 (#48)	~80%	~91%	—	High
Grok 4.3	US	38 (high)	1443 (#59)	87,5% (Grok 4)	~94%	~79% (LiveCodeBench)	High
Claude Haiku 4.5	US	30	1411 (#107)	~73%	—	73,3%	High
Gemini 2.5 Pro	US	27	1450	84–86%	88,0%	63,8%	High
ERNIE 5.0	CN	22 (think)	1447 (#52)	~73% (self)	~80% (self)	—	Med
Llama 4 Maverick	US	14,3	1417	69,8%	—	—	High
Gemini 2.5 Flash	US	14,1	1410 (#109)	~78%	~72%	—	High

> *Контекст шкалы: #1 рынка — **Claude Fable 5** (AA 60, Elo 1508). Разрыв «топ США ↔ топ Китая» по AA сжался до Fable 5 (60) → Opus 4.8 (56) → GLM-5.2 / Qwen3.7-Max (51 / 46) — китайский фронтир отстаёт примерно на один релиз, а не на поколение.*

Россия — отдельная шкала (нет AA/Arena), данные по MERA — самооценка вендоров

Модель	MERA (RU композит)	MMLU (EN)	ruMMLU	MMLU-Pro	Примечание	Conf
GigaChat 2 Max	**0,67**	0,86	0,805	0,667	По статье Сбера (arXiv 2506.09440) обходит GPT-4o (0,642) на MERA	Med (vendor)
GigaChat 2 Pro	0,649	0,821	0,775	0,644		Med (vendor)
YandexGPT 5.1 Pro	нет публичного MERA	—	—	—	Самооценка: бьёт GPT-4.1 в ~56% слепых человеческих сравнений	Low (vendor)

Инсайт по качеству. (1) Разрыв «топ США ↔ топ Китая» по AA Index сократился до ~5 баллов (Opus 4.8 = 56 / GLM-5.2 = 51) — это уже не пропасть, а догоняющая дистанция в один релиз. (2) LMArena (человеческие предпочтения) и AA Index (способность) **расходятся**: Gemini 3.1 Pro в Arena стоит выше (1486, #7), чем по AA (46), тогда как GLM-5.2 наоборот сильнее по AA (51), чем по Arena (1470, #25) — «нравится людям» и «решает бенчмарки» — это две разные оси, и наш композит 0,6/0,4 их намеренно смешивает. (3) Внутри Anthropic интересна вилка: **Fable 5** (AA 60, Elo 1508) ощутимо выше Opus 4.8 (56 / 1484), но и вдвое дороже ($10/$50 против $5/$25). (4) Российские модели нельзя сравнивать с мировыми напрямую: MERA — русско-специфичный бенчмарк, цифры — самооценка вендоров и не воспроизведены на западных харнессах. Принимать как «GigaChat ≈ GPT-4o» некорректно — это уровень *на русскоязычных задачах по версии Сбера*.

---

6. Сводный параметр: рейтинги «качество», «цена», «интеллект на доллар»

Методология — в разделе «Введение». Ниже три ранжирования по 31 модели с полными данными (RU — отдельно, без AA/Arena).

6.1. Топ по QUALITY (композит AA + Arena, 0–100)

#	Модель	Регион	Quality	Blended $	Value
1	Claude Fable 5	US	**97,8**	20,00	4,9
2	Claude Opus 4.8	US	89,4	10,00	8,9
3	GPT-5.5	US	87,9	11,25	7,8
4	GPT-5.4	US	83,4	5,63	14,8
5	Gemini 3.5 Flash	US	82,0	3,38	24,3
6	GLM-5.2	CN	81,9	2,15	38,1
7	Gemini 3.1 Pro	US	79,8	4,50	17,7
8	Claude Sonnet 4.6	US	78,3	6,00	13,0
9	Qwen3.7-Max	CN	77,8	0,52	**149,0**
10	DeepSeek-V4-Pro	CN	72,5	0,54	133,3

6.2. Топ по VALUE (интеллект на доллар = Quality ÷ blended)

#	Модель	Регион	Value	Quality	Blended $
1	DeepSeek-V4-Flash	CN	**369**	64,6	0,175
2	Qwen-Flash	CN	351	20,0	0,057
3	MiniMax M3	CN	160	70,7	0,44
4	Qwen3.7-Max	CN	149	77,8	0,52
5	Llama 4 Maverick	US	136	35,6	0,26
6	DeepSeek-V4-Pro	CN	133	72,5	0,54
7	Doubao-Seed-1.6	CN	120	43,3	0,36
8	Gemini 2.5 Flash-Lite	US	95	16,7	0,175
9	ERNIE X1	CN	76	36,7	0,485
10	Qwen-Plus	CN	67	40,0	0,60

> ***8 из 10 строк топ-Value — Китай.*** *Неазиаты в десятке — только открытая Llama 4 Maverick (Meta, сторонний инференс) и Gemini 2.5 Flash-Lite. Абсолютный лидер value —* ***DeepSeek-V4-Flash****: 65 баллов качества (уровень среднего фронтира) за blended $0,175.*

6.3. Топ по ЦЕНЕ (blended, дешевле — выше)

#	Модель	Регион	Blended $	Вход / Выход
1	Qwen-Flash	CN	0,057	0,033 / 0,13
2	Hunyuan-TurboS	CN	0,153	0,11 / 0,28
3	DeepSeek-V4-Flash	CN	0,175	0,14 / 0,28
3	Gemini 2.5 Flash-Lite	US	0,175	0,10 / 0,40
5	Llama 4 Maverick	US	0,262	0,15 / 0,60
6	Doubao-Seed-1.6	CN	0,36	0,11 / 1,11
7	GLM-4.5-Air	CN	0,425	0,20 / 1,10
8	MiniMax M2	CN	0,441	0,255 / 1,00
9	GPT-5.4-nano	US	0,463	0,20 / 1,25
10	DeepSeek-V4-Pro	CN	0,544	0,435 / 0,87

Главный инсайт сводного параметра. Качество и цена слабо коррелируют поперёк границ. Внутри США премия за качество почти линейна и крута. Но **китайские модели сломали эту кривую**: DeepSeek-V4-Pro даёт ~74% от качества абсолютного лидера (Fable 5) за 2,7% его цены; Qwen3.7-Max — 80% качества при value 149. Для подавляющего большинства текстовых задач (где не нужны последние 15–20 баллов интеллекта) рациональный выбор по экономике — китайский frontier-second-tier (DeepSeek V4, GLM-5.2, Qwen3.7, MiniMax M3), а топ-флагманы США (Fable 5, Opus 4.8, GPT-5.5) оправданы там, где маржинальные баллы качества стоят денег: юридический/медицинский текст, сложное рассуждение, агентные цепочки с накоплением ошибки. Отдельно отметим **Gemini 3.5 Flash** — единственный западный, попавший в топ-5 качества (82) при умеренной цене ($3,38, value 24): лучший западный баланс «цена/качество» на фронтире.

---

7. Подписки и включённый объём токенов

Запрос требовал учесть подписки «по объёму токенов, включённых в подписку». **Критическая находка: почти никто не публикует токенные квоты.** ChatGPT, Claude, Gemini в 2026 перешли на «скользящее окно (3–5 ч) + недельный лимит» и раскрывают только лимиты *сообщений*, не токенов. Все токен-оценки ниже — расчётные (потолок = cap × число окон × ~1500 ток/сообщение), реальное использование обычно 10–25% потолка. Помечены **ESTIMATE**.

США

Сервис	Тариф	Цена/мес	Модели	Лимит (публичный)	Токенов/мес (оценка)	Conf
ChatGPT	Free	$0	GPT-5.x Instant	~10 сообщ/5ч	~2M (ESTIMATE)	Low
ChatGPT	Plus	$20	GPT-5.5 + Thinking	160 сообщ/3ч + 3000 Thinking/нед	~55M потолок (ESTIMATE)	Low
ChatGPT	Pro	$200	Все, Thinking почти-безлимит, 400K ctx	Без жёсткого cap	практически безлимит	High (no cap)
Claude	Free	$0	Sonnet 4.6	~25 сообщ/5ч + нед. cap	~3–5M (ESTIMATE)	Low
Claude	Pro	$20	Sonnet 4.6 + Opus 4.8	~45 сообщ/5ч + нед. cap	~9–10M (ESTIMATE)	Low
Claude	Max 5×	$100	Sonnet + Opus	~88K ток/5ч	~12–13M (ESTIMATE)	Low
Claude	Max 20×	$200	Sonnet + Opus	~220K ток/5ч	~30–32M (ESTIMATE)	Low
Gemini	AI Plus	$8	Gemini 3.x	2× «стандарта»	~2–6M (ESTIMATE)	Low
Gemini	AI Pro	$20	Gemini 3.1 Pro, 1M ctx	4× «стандарта»	~4–12M (ESTIMATE)	Low
Gemini	AI Ultra	$200	3.1 Pro / Deep Think	20× AI Pro	~80–240M (ESTIMATE)	Low
Grok	SuperGrok	$30	Grok 4.x	текст щедро; 50 DeepSearch/день	~10–30M (ESTIMATE)	Low
Grok	SuperGrok Heavy	$300	Grok 4.3 full	200 DeepSearch/день	50M+ (ESTIMATE)	Low
Perplexity	Pro	$20	GPT-5.4, Opus, Gemini 3.1	Unlimited Pro Search; 20 Deep Research/день	~10–30M (ESTIMATE)	Low

Китай

Сервис	Тариф	Цена/мес	Модели	Включено	Conf
DeepSeek	Free app	¥0	V4-Flash + V4-Pro, DeepThink	Без публичного cap (best-effort); API: 5M триал-токенов	High (free)
Doubao	Standard / Enhanced / Pro	¥68 / ¥200 / ¥500 (~$9,5/$28/$70)	Doubao	Квоты по задачам, токены не публикуются	Low
Kimi	Moderato / Allegretto / Allegro / Vivace	$19 / $39 / $99 / $199	K2.5/K2.6, 256K ctx	Квоты, токены не публикуются	Low
Qwen	Free app	¥0	Qwen Chat	Без cap для обычного использования	High (free)
Ernie Bot	Free / Premium	¥0 / ¥59,9 (~$8,3)	ERNIE 4.5 + X1	Бесплатен с 04.2025; Premium — расширенные лимиты	High/Low

Россия

Сервис	Тариф	Цена/мес	Включено	Conf
Yandex (Алиса AI)	Free	0₽	5 запросов/день; Alice Pro free: 50 запросов/мес	Low (ESTIMATE)
Yandex	Алиса Плюс	+100₽ (~$1,25)	«Безлимит» reasoning-ответов в Поиске с Алисой	Low
Yandex	Яндекс Плюс	399→449₽ (~$5–5,6)	YandexGPT Pro через Alice Pro, расширенные лимиты	Low
Sber	**GigaChat Freemium**	0₽	**1 000 000 токенов / 12 мес (~83K/мес)** — единственный токен-номинированный потреб-тариф	**High**
Sber	GigaChat Premium	от 290₽ (~$3,6)	GigaChat 2 Pro + Max; pay-as-you-go 0,2₽/1K	Low
T-Bank	Gen-T assistants	0₽	Бесплатно в приложении, лимиты не публикуются	High (free)

Инсайт по подпискам. (1) Перевод подписок на «окна + недельные лимиты» — это намеренный отказ от токенной прозрачности: провайдер управляет нагрузкой, не давая пользователю считать «себестоимость токена». (2) Для тяжёлого пользователя подписка почти всегда дешевле API: Claude Max 20× за $200 при ~30M токенов/мес эквивалентен ~$6,7/1M blended на Opus — это *дешевле* API-ставки Opus ($10 blended), плюс безлимит в окне. (3) Китайские потребительские приложения

(DeepSeek, Qwen, Ernie) фактически **бесплатны без явных лимитов** — субсидируемая гонка за долю рынка. (4) GigaChat Freemium — редкий честный токенный грант, удобен для пилотов/MVP.

---

8. Синтез и стратегические выводы

Три рынка, три логики ценообразования.

- **США = премия за потолок.** Цена растёт круто и почти линейно с качеством; флагманы продают «последние 10–15 баллов» с премией ×3–5. Экосистема, инструменты, надёжность, агентные возможности — оправдывают премию для бизнес-критичных задач.
- **Китай = демпинг за долю рынка.** Кэш-скидки до –98%, batch –50%, бесплатные потребительские приложения, frontier-качество за 5–20% цены США. Это сознательная стратегия захвата доли — экономика, вероятно, отрицательная. Риск: доступность/санкции/блокировки, материковый vs международный эндпойнт, переменчивость промо-цен.
- **Россия = суверенитет, а не цена.** RU-модели не дешевле и не сильнее мировых; их ценность — 152-ФЗ, ПД в РФ, рубли, отсутствие санкционных рисков, экосистемная интеграция. Для регулируемых отраслей (госсектор, банки, медицина) это безальтернативно.

Что коррелирует, а что нет. Цена и качество сильно коррелируют *внутри* одного провайдера и слабо — *поперёк границ*. Это арбитраж: одну и ту же текстовую задачу можно решить за $0,5 (китайский второй эшелон) или за $11 (флагман США) с разницей качества в ~20%. Рациональная стратегия — **роутинг по сложности**: дешёвая модель на массовый объём (классификация, извлечение, черновики), флагман — только на сложное рассуждение.

Reasoning размывает токенную экономику. Рассуждающие модели (o-серия, R1/V4-Pro, Thinking-режимы) тратят в 3–15× больше выходных токенов на «размышление». Номинальная цена за токен перестаёт отражать стоимость задачи — надо считать стоимость *решённой задачи*, а не токена. Новый токенизатор Opus 4.7+ (+35% токенов) — пример того же эффекта со стороны входа.

---

9. Ограничения и оговорки

1. **Снимок, а не тренд.** Цены LLM меняются ежемесячно, лидерборды — еженедельно. Всё датировано 26.06.2026; через квартал значимая часть цифр устареет.
2. **Confidence неоднороден.** Официальны (High): Anthropic, OpenAI (5.x), Google, Mistral, DeepSeek, Zhipu, GigaChat, YandexGPT-флагман. Агрегаторы/пресса (Med/Low): Amazon Nova, Llama, Qwen/Doubao/ERNIE/Hunyuan в CNY, Kimi/MiniMax, MTS Cotype, страница тарифов Яндекса (CAPTCHA).
3. **Leaderboard-данные прочитаны живьём, но это снимок одного дня.** AA Index и Elo для всех топ-моделей считаны из рендеренного DOM artificialanalysis.ai и lmarena.ai 26.06.2026 (Playwright, не оценка) — High. Но: (а) AA даёт несколько вариантов по «усилию рассуждения» — мы брали старший стандартный, другой выбор сдвинул бы баллы; (б) у плотного топа Arena CI перекрываются, ранги ±несколько позиций; (в) для GLM-4.7, Doubao-Seed-1.6, MiniMax M2 точного

варианта на бордах нет — их AA взяты по ближайшему релизу/оценке (Med/Low). Composite надёжен на ±5 баллов.

4. **AA Index пересмотрен** в начале 2026 — значения не сравнивать с «launch-era ~60–70» из старых обзоров.
5. **Российские бенчмарки — самооценка вендоров** (MERA по статьям Сбера, человеческие SBS Яндекса), не воспроизведены независимо. Прямое «GigaChat ≈ GPT-4o» некорректно.
6. **Подписочные токены — расчётные потолки.** Ни один из ChatGPT/Claude/Gemini не публикует месячных токенных квот; реальное потребление — 10–25% от оценок.
7. **Ступенчатые цены и кэш** не отражены в blended полностью: реальная стоимость в продакшене с кэшем/батчем у китайских моделей ниже номинала, у длинно-контекстных (Gemini Pro >200K) — выше.

---

10. Рекомендации: какую модель под какую задачу

Сценарий	Рекомендация	Почему
Абсолютный максимум интеллекта, цена не важна	**Claude Fable 5**	#1 и в LMArena (1508), и в AA Index (60)
Максимум интеллекта при разумной премии	**Claude Opus 4.8** / GPT-5.5	#2–3 по качеству, вдвое дешевле Fable 5
Та же лига качества, но с оглядкой на бюджет	**GLM-5.2** ($2,15) / **DeepSeek-V4-Pro** ($0,54)	74–82% качества лидера за 3–11% цены
Лучший баланс среди западных	**Gemini 3.5 Flash** ($3,38) / Grok 4.3 ($1,56)	Quality 64–82 при умеренной цене
Массовый дешёвый текст (классификация, извлечение, черновики)	**DeepSeek-V4-Flash**, **Qwen-Flash**, **Gemini 2.5 Flash-Lite**	$0,06–0,18 blended при достаточном качестве
Кодинг	GPT-5.3-codex, Grok 4.3, Kimi K2.7 Code, DeepSeek-V4-Pro	Профильная заточка / высокий SWE-bench
Данные обязаны остаться в РФ (152-ФЗ)	**GigaChat 2 Max/Pro** или **YandexGPT 5.1 Pro**	Локализация, рубли, отсутствие санкц-рисков; GigaChat дешевле
Пилот/MVP без бюджета	GigaChat Freemium (1M ток), DeepSeek/Qwen free app, GLM-4.7-Flash (free)	Бесплатный вход
Тяжёлый личный пользователь (chat)	Claude Max 20× / ChatGPT Pro	Эффективная ставка < API при безлимите в окне
Открытые веса / self-host	Llama 4, T-Pro 2.0 (RU), GLM/Qwen open releases	Контроль, приватность, нулевая цена за токен

Библиография

США — официальные тарифы: OpenAI (developers.openai.com/api/docs/pricing; openai.com/api/pricing) · Anthropic (platform.claude.com/docs/en/about-claude/pricing) · Google Gemini (ai.google.dev/gemini-api/docs/pricing) · xAI (docs.x.ai/docs/models) · Mistral (mistral.ai/pricing) · Cohere (cohere.com/pricing) · Amazon Bedrock/Nova (aws.amazon.com/bedrock/pricing) · Azure Phi (azure.microsoft.com/pricing/details/ai-foundry-models).

Китай — официальные/агрегаторы: DeepSeek (api-docs.deepseek.com/quick_start/pricing) · Zhipu z.ai (docs.z.ai/guides/overview/pricing; bigmodel.cn/pricing) · Alibaba Qwen (help.aliyun.com/zh/model-studio/model-pricing) · Moonshot Kimi (platform.kimi.ai/docs/pricing) · Baidu Qianfan (cloud.baidu.com/product-s/qianfan_home) · ByteDance Volcengine (volcengine.com/docs/82379) · MiniMax (platform.minimax.io/docs/guides/pricing-paygo) · Tencent Hunyuan (cloud.tencent.com/document/product/1729).

Россия: Yandex (пресс-релиз 28-08-2025-01; ai-journal.ru/yandexgpt; aistudio.yandex.ru) · Сбер GigaChat (developers.sber.ru/docs/ru/gigachat/tariffs) · T-Bank (huggingface.co/t-tech; habr.com/companies/tbank) · MTS Cotype (tadviser; mts.ai/product/cotype).

Бенчмарки и лидерборды: Artificial Analysis (artificialanalysis.ai/leaderboards/models; /evaluations/aime-2025) · LMArena (lmarena.ai; llm-stats.com/benchmarks/lmarena-text; huggingface.co/spaces/lmarena-ai) · GPQA/MMLU-Pro (llm-stats.com; pricepertoken.com) · Coding (morphllm.com/swe-bench-pro) · Китайские модели (benchlm.ai/blog/best-chinese-llm; turingpost.com/p/chinesemodels; GLM-4.5 arXiv 2508.06471) · Россия (GigaChat arXiv 2506.09440; mera.a-ai.ru; mysummit.school/yandexgpt-review-2026) · Кросс-чек (vellum.ai/llm-leaderboard; clickrank.ai/llm-leaderboard).

Подписки: OpenAI Help (help.openai.com/articles/11909943) · Claude Help (support.claude.com/articles/11647753) · TokenMix · Google (blog.google/.../google-ai-subscriptions; support.google.com/gemini/answer/16275805) · xAI (felloai.com/grok-pricing) · Perplexity (finout.io/blog/perplexity-pricing-in-2026) · DeepSeek (datastudios.org) · Doubao (caixinglobal.com 2026-05-05) · Kimi (kimi.com/membership/pricing) · Baidu (prnewswire.com) · Yandex (alicepro.yandex.ru; yandex.ru/support/plus-ru) · GigaChat (developers.sber.ru/.../individual-tariffs).

Приложение: методология сводного параметра

Источник данных качества. Две оси: (1) Artificial Analysis Intelligence Index — композит MMLU-Pro + GPQA Diamond + AIME + LiveCodeBench + др., шкала июня-2026 (топ ~56); (2) LMArena text Elo — человеческие предпочтения в слепых дуэлях.

Нормировка. `Q_aa = AA/56·100`; `Q_arena = clamp((Elo−1300)/220·100, 0, 100)`. `Quality = 0,6·Q_aa + 0,4·Q_arena` (если есть обе оси; иначе доступная). Вес 0,6 в пользу способности — оптимизация под «работу с текстом».

Цена. `blended = (3·вход + 1·выход)/4` — конвенция Artificial Analysis (типичное реальное соотношение 3:1 вход:выход). Кэш и ступени контекста в blended не учтены (см. ограничение 7).

Value. `Value = Quality / blended` — «единиц качества на $1 за 1M blended-токенов». Безразмерный, сравним только внутри этой выборки и снимка.

Воспроизводимость. Полный датасет и расчёт — `compute_scores.py` и `scores.json` в папке отчёта. RU-модели исключены из композита (нет AA/Arena) и разобраны по MERA отдельно.

Отчёт подготовлен 26.06.2026. Все цифры — снимок на эту дату; цены LLM меняются ежемесячно, лидерборды — еженедельно. Перепроверяйте перед использованием в решениях.