Воркшит AI Video Cost Model

Подставьте нагрузку в модель «стоимость на час видео». Цены мая 2026.

1. Цены закрытых API (USD за 1M токенов, если не указано иначе)

Модель	Input $/1M	Output $/1M	Cache read	Где использовать
Gemini 2.5 Flash-Lite	0.10	0.40	0.01	Классификация, роутинг
Gemini 2.5 Flash	0.30	2.50	0.03	Дефолт видео-нагрузок
Gemini 2.5 Pro	1.25	10.00	0.125	Long-context видео
GPT-5	0.625	5.00	varies	Рабочее рассуждение
GPT-5.5	5.00	30.00	varies	Flagship-рассуждение
Claude Haiku 4.5	1.00	5.00	0.10	Быстрый структурный вывод
Claude Sonnet 4.6	3.00	15.00	0.30	Long-form, агенты
Claude Opus 4.7	5.00	25.00	0.50	Самые сложные задачи
OpenAI Realtime audio	32	64	0.40	Realtime-голос
Gemini Live audio	3.00	12.00	varies	Realtime голос/видео

2. Аудио-AI и генеративное видео — стоимость на час контента

Аудио (ASR) на час обработки
Deepgram Nova-3 batch: $0.26
Deepgram Nova-3 streaming: $0.46
OpenAI gpt-4o-mini-transcribe: $0.18
OpenAI gpt-4o-transcribe: $0.36
AssemblyAI Universal Streaming: $0.15
Azure Speech real-time: $1.00
Google STT V2 base: $0.96
Self-host Whisper (только compute): ~$0.03

Генеративное видео на час вывода
Hailuo 02 Standard 768p: $162
Runway Gen-4 Turbo: $180
Veo 3.1 Lite 720p: $180
Kling 3.0 720p: $270
Sora 2 Standard 720p: $360
Sora 2 Pro 720p: $1 080
Sora 2 Pro 1024p: $1 800
Veo 3 с нативным аудио: $2 700

3. Пример — видеоконференц-продукт на 100 000 MAU

Нагрузка (дешёвый билд)	Объём в мес	Стоимость в мес
Live-субтитры (AssemblyAI Universal Streaming)	800 000 ч	$120 000
Real-time перевод (Gemini Live, 20% звонков)	160 000 ч	$172 800
AI-нотетейкер (Gemini 2.5 Flash)	800 000 звонков	$5 120
Еженедельные дайджесты (Gemini 2.5 Flash)	400 000 дайджестов	$8 450
Векторное хранение (pgvector self-host)	1.8B embeddings	$2 000
Compliance + ops overhead (множитель 1.5×)	—	~$45 000
Итого дешёвый билд	—	~$353 000
(Наивный билд для сравнения)	—	~$700 000

4. Рычаги стоимости — стакайте все

Роутинг на дешевле (×5–×30) · Prompt caching (70–90% на cached) · Batch (50%) · Low-resolution media (×3 на видео) · Context truncation (20–60%) · System-prompt compression (10–30%) · Self-host на open weights (×5–×20) · Edge inference (до ×100) · Async-пайплайны (50–80%) · Volume-контракты (20–40% выше $50K/мес) · Cost-per-request мониторинг (10–30%) · Spot GPU (50–80%).