

Чек-лист выбора движка потокового TTS

Сравнение задержки и цены 2026, хости-или-арендуй и пре-флайт перед запуском голосовой фичи

1 · Варианты потокового TTS 2026 с одного взгляда

Движок	Хост / аренда	TTFA P50	Цена	Языки	Лучше всего для
Kokoro-82M	Self-host	—*	~\$1 / М симв.	8	Объём, приватность, пол цены
ElevenLabs Turbo v2.5	Хостинг API	264 мс	\$0.10 / 1k	32	Языки + стабильность
ElevenLabs Flash v2.5	Хостинг API	288 мс	\$0.05 / 1k	32	Дёшево + низкая задержка
Cartesia Sonic-3	Хостинг API	188 мс	от \$0	40+	Быстрейшая медиана
OpenAI gpt-4o-mini-tts	Хостинг API	поток	~\$0.015/мин	мульти	Пакетная озвучка в OpenAI
OpenAI TTS-1-HD	Хостинг API	2 295 мс	\$30 / М	мульти	Только пакетно, не realtime

TTFA = медианное time-to-first-audio (Coval, май 2026). *Задержка Kokoro зависит от железа (~210× realtime на одной RTX 4090). Цельтесь в < 300 мс TTFA; мерьте TTFA, не time-to-first-byte.

2 · Хости сам или арендуй

Критерий	Хости сам (Kokoro)	Арендуй (ElevenLabs / Cartesia / OpenAI)
Модель цены	GPU + электричество (почти ноль)	За символ текста
Настройка	Вы держите и масштабируете серверы	Старт за полдня
Приватность	Аудио остаётся в вашей сети	Аудио уходит к вендору
Языки	8	32 - 40+
Лучше на объёме	Очень большой объём	Малый-средний объём

3 · Пре-флайт перед запуском потоковой голосовой фичи

- ☐ Определили реальное время или пакет — если вживую никто не ждёт, выбирайте по качеству/цене; если разговорно, задержка правит всем.
- ☐ Задали бюджет задержки: цель < 300 мс TTFA (ниже 200 мс — отлично), помня, что TTS идёт после STT + LLM в цепочке.
- ☐ Меряли time-to-first-audio, не time-to-first-byte — отбросьте заголовок файла перед замером первого звука.
- ☐ Взвесили стабильность задержки (разброс / IQR), не только медиану — широкий разброс ломает долю вызовов на масштабе.
- ☐ Посчитали месячный объём в символах (~900 на минуту речи) против ставки каждого движка за символ.
- ☐ Смоделировали безубыточность самохостинга: на больших объёмах почти нулевая стоимость Кокоро бьёт счёт за символ.
- ☐ Подтвердили нужное покрытие языков (Kokoro 8 против ElevenLabs 32 против Cartesia 40+).
- ☐ Проверили ограничение приватности: если аудио не может покидать сеть — хостите open-модель сами.
- ☐ Выбрали верный транспорт: SSE, когда весь текст есть заранее, переиспользуемый WebSocket, когда текст идёт от LLM.
- ☐ Настроили chunk schedule на реальном выводе LLM и держали один WebSocket открытым на сессию (экономия 50-100 мс/ход).
- ☐ Выбрали серверный регион рядом с пользователями (или хостили близко к ним), чтобы срезать сетевой путь.
- ☐ Проверили худший случай, не только демо: тестировали сквозной TTFA из реальных локаций пользователей.