Памятка: выбор архитектуры speech-to-speech

Каскад против end-to-end, бюджет 800 мс, выбор транспорта, смена реплик и сравнение систем 2026

1 · Три системы 2026 одним взглядом

Система	Задача	Держать / аренда	Подключение	Стоимость	Когда брать
OpenAI Realtime API	Разговор	Аренда (API)	WebRTC/WS/SIP	~$0,18-0,46/мин	Продакшен-агент, телефон
Gemini Live	Разговор	Аренда (API)	WebSocket	~10× дешевле вход	Высокий объём, бюджет
SeamlessM4T v2	Перевод	Держать (open)	Связываете сами	GPU + электр.	Открытая база перевода

OpenAI и Gemini — end-to-end-движки разговора; SeamlessM4T — открытая модель перевода под лицензией CC-BY-NC (некоммерческой) — проверьте лицензию перед выпуском. Цены за минуту и за токен прочитаны в мае 2026; перепроверяйте при использовании.

2 · Каскад или end-to-end?

Критерий	Каскад (3 модели)	End-to-end (1 модель)
Как работает	Цепочка ASR -> LLM -> TTS, текст между этапами	Одна модель: аудио вход, аудио выход, без середины
Задержка	Складывается из 3 передач; растёт с моделью	Ниже — передач нет
Отлаживаемость	Высокая — текст на каждом этапе	Низкая — нет читаемой середины
Естественность	Голос уплощён в текст и обратно	Сохраняет смешки, интонацию, перебивания
Привязка	Заменить любой компонент, любой вендор	Один вендор на все три задачи
Когда брать	Нужны контроль и отлаживаемость	Разговор должен быть живым и быстрым

3 · Перед запуском speech-to-speech-фичи

- ☐ Назвали задачу: разговор (ответ на том же языке) или перевод (на другом языке)?
- ☐ Выбрали архитектуру: каскад для контроля/отладки, end-to-end для низкой задержки и человечности.
- ☐ Задали бюджет от голоса до голоса: цель < 800 мс; модель — лишь ~половина (≈ 500 мс до первого байта в США).
- ☐ Измерили сквозную задержку голос-в-голос, а не только время ответа модели, из реальных локаций пользователей.
- ☐ Выбрали транспорт: WebRTC для аудио с устройства, WebSocket для серверных пайплайнов, SIP для телефонных звонков.
- ☐ Настроили определение конца реплики (VAD) на реальном «грязном» аудио — акценты, шум, паузы — не в тихом демо.
- ☐ Сделали обработку перебиваний (barge-in) полноценной фичей: остановить вывод, очистить очередь, перезапустить слушание.
- ☐ Прикинули минуты/месяц против прайсов и OpenAI, и Gemini, включая скидки за prompt caching.
- ☐ Включили prompt caching, где доступно (OpenAI кэширует аудиовход ~в 80× дешевле свежего).
- ☐ Подтвердили охват языков для нужных вам языков, включая переключение посреди фразы, если нужно.
- ☐ Проверили лицензию при своём хостинге: SeamlessM4T — CC-BY-NC, коммерческое использование требует отдельной договорённости.
- ☐ Проверили место хранения данных и приватность: куда едет аудио и отвечает ли это вашим требованиям?