

LiveKit AI Meeting Assistant — Cheat Sheet

Fora Soft Learn · AI for Video Engineering

1 - How an AI agent joins a meeting (it is an extra participant)

User app -> Your backend (room + token, dispatch)
 -> LiveKit server (Cloud / self-hosted)
 -> AgentServer launches an isolated job -> Agent joins the room as a participant.
 Agent loop: subscribe to each audio track -> STT -> LLM -> (optional) TTS
 -> publish transcript / recording / spoken audio back into the room.
 Dispatch < 150 ms. Each job is its own process; it runs until everyone leaves the room.

2 - Three voice-pipeline types (choose by what the feature needs)

STT -> LLM -> TTS : live transcripts, most control. Default for a notetaker.
 Realtime model : most natural, but NO interim transcript - add STT for captions.
 Half-cascade : realtime understanding + a separate TTS voice. A blend of both.

3 - LiveKit Cloud tiers (2026 - re-verify at livekit.io/pricing)

Build \$0/mo	5,000 WebRTC min, 1,000 agent min, 50 GB, 5 concurrent
Ship \$50/mo	150k WebRTC (\$.0005), 5k agent (\$.01), 250 GB (\$.12)
Scale \$500/mo	1.5M WebRTC (\$.0004), 50k agent (\$.01), 3 TB (\$.10)
Enterprise	Custom pricing and concurrency.
Self-host	Open-source server + Agents = free (Apache 2.0); pay own infra.
Note	Model costs (STT / LLM / TTS) are billed on top of every tier.

4 - What a minute costs (representative 2026 model rates)

Listen-only notetaker = agent \$0.01 + STT \$0.006 + observability \$0.01 ~ \$0.026 / min
 Talking voice agent = + LLM \$0.0077 + TTS \$0.030 ~ \$0.0635 / min
 Text-to-speech is the heaviest layer; dropping it roughly halves the per-minute cost.

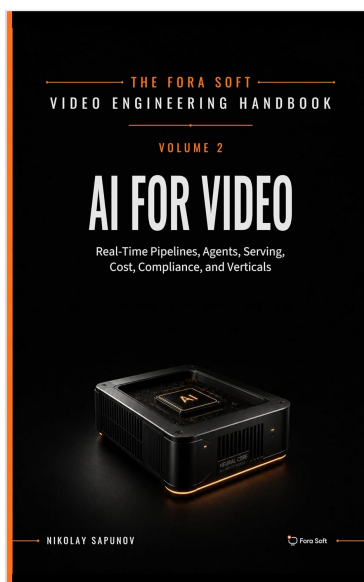
5 - Build-vs-buy checklist

- ☐ Decide first: is the assistant your product, or a commodity transcript feature?
- ☐ Need live transcripts? Use STT-LLM-TTS, or add STT alongside a realtime model.
- ☐ Listen-only notetaker? Drop text-to-speech to roughly halve the per-minute cost.
- ☐ Strict privacy or data residency? Self-host the open-source server (Apache 2.0).
- ☐ Just want transcripts fast? Buy a meeting-bot service (e.g. Recall.ai ~\$0.50/hr).

Companion to: 'LiveKit Real-Time AI Meeting Assistant — Architecture and Pricing'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/livekit-realtime-ai-meeting-assistant-news-pricing

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 7 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Real-Time Pipelines and Agents to Serving, Cost, Compliance, and Verticals

By Nikolay Sapunov, CEO at Fora Soft · 1,221 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0GX3DGBH4

This cheat sheet shows how a LiveKit agent joins a meeting, the three voice-pipeline types and what a minute costs. Volume 7 covers this in depth: latency budgets for calls, where AI enters the WebRTC pipeline, agent loop primitives and memory, serving infrastructure, and cost per stream-hour.

WHAT YOU WILL FIND INSIDE

- Latency budgets for AI in calls
- Agent loops, memory and tool use
- Serving with vLLM, Triton and TensorRT
- Cost per stream-hour and how to cut it

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Building an AI meeting assistant on LiveKit or another WebRTC stack?

We will pick the voice pipeline, models and hosting for your assistant and estimate the cost per meeting minute. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)