Real-Time AI Latency Budget Worksheet

Fora Soft Learn · AI for Video Engineering

1 - The target (ITU-T G.114 mouth-to-ear delay)

```
< 150 ms = transparent   |   150-400 ms = usable but degraded   |   > 400 ms = unacceptable
Design conversational AI to < 200 ms (natural turn gap); aim for 100 ms to keep headroom.
```

2 - Fixed pipeline (you do NOT control these - sum them first)

Capture (camera + mic)	3-10 ms
Video encode / decode (hardware, each)	~10 ms
Audio codec - Opus, RFC 6716 (default)	26.5 ms
Network propagation	~10 ms per 1,000 km RTT
TURN relay / SFU forwarding	10-30 / 5-20 ms
Jitter buffer (real-time, adaptive)	10-50 ms
Render to display (60 Hz)	8-16 ms

3 - AI inference reference numbers (2026 - fast-moving)

On-device segmentation - MediaPipe (GPU)	< 3 ms / frame
Streaming ASR first words - Deepgram Nova-3	~150 ms interim, < 300 ms
Streaming TTS first audio - ElevenLabs Flash / Cartesia Sonic	~75-90 ms
Voice-to-voice first audio - OpenAI Realtime	~300-500 ms (does NOT fit a live call)

4 - The speed-of-light tax (geography, not model speed)

```
Light in fiber ~ 200,000 km/s -> ~10 ms RTT per 1,000 km. London <-> N. Virginia ~ 60 ms RTT
before any router. A distant cloud GPU can spend your whole budget on the network alone.
```

5 - Placement rule + the back-of-envelope sum

- [ ] Compute the fixed sum FIRST. Regional call example: 60 (net) + 26.5 (Opus) + 40 (jitter) + ~35 (capture/encode/decode/render) = ~161 ms.
- [ ] Budget left for AI = target - fixed sum. Against 200 ms that is ~39 ms. The model must fit THAT, or the feature is not live.
- [ ] Instant + small model -> run ON-DEVICE (zero network). Live + big model -> run at the EDGE (< 20 ms).
- [ ] Tolerates 1-3 s -> run in the CLOUD (summaries, analytics, copilots). Do not fight physics for these.
- [ ] Benchmark with the network distance your real users have - never next to the server.