

CODE IS THE FLOOR. STANDARD IS THE CEILING.

Scope: Operator-grade local AI for luxury residential boardrooms and home offices. The room hears. The room understands. Nothing leaves the room. Mac Studio M2 Ultra · Llama 3.3 70B · Whisper · Piper · Ollama · Sennheiser TCC 2 · Shure MXA920 · Crestron · Lutron · BGE-M3 · QC.

§01	OPERATING PRINCIPLES
- Local-only inference — model weights, transcripts, and embeddings never leave the room. - Air-gapped at the WAN: AI VLAN egress denied; firmware updates pulled by hand on schedule. - Operator-grade hardware tier matched to data sensitivity — boardroom, family office, or add-on. - AI captures, drafts, and answers. The principal decides. The operator is accountable. - Every spoken control intent passes a confirmation step — no autonomous commits to live systems. "AI says so" is not authority. Verification protocol applies before any field decision.	

§02	HARDWARE TIER MATRIX
- Tier 1 Boardroom: Mac Studio M2 Ultra 192 GB · Llama 3.3 70B Q4 · 6 dBA idle · 3U Sonnet. - Tier 1 alt: NVIDIA RTX 4090 24 GB in 2U workstation — sub-20B models, plan thermal isolation. - Tier 2 Family Office: Mac Mini M4 Pro 64 GB · Llama 3.1 8B + Whisper Large + Piper TTS. - Tier 3 Add-On: Mac Mini M4 base 16 GB — transcription-only, no LLM, no RAG. - Quantization standard: Q4_K_M production · Q5_K_M Tier 1 quality · Q8_0 embeddings. - Throughput: M2 Ultra 70B Q4 ≈ 20-30 tok/s · M4 Pro 8B Q4 ≈ 40-60 tok/s · sub-2 s response.	

§03	SOFTWARE STACK
- Ollama — model server, REST on 127.0.0.1:11434, OpenAI-compatible at /v1; no auth, localhost-bound. - Open WebUI — browser UI, multi-user sessions, document upload, Docker volume only; no telemetry. - Whisper.cpp — Core ML build for Apple Silicon · large-v3 · 3x faster than real-time on Neural Engine. - faster-whisper — CTranslate2 build for RTX 4090 Tier 1 · int8 batch_size=8 ≈ 16 s on 13 min audio. - Piper TTS — local neural voice · GPL-3.0 (OHF-Voice piper1-gpl) · routed to room DSP. - llama.cpp foundation · vLLM for multi-user Tier 1 GPU · LM Studio approved for Tier 2 desktop.	

§04	WITHOUT WRITTEN DIRECTION
- We do not run AI in the conference room without written direction: cloud transcription of any audio. - We do not run AI in the conference room without written direction: PII through any consumer AI tier. - We do not run AI in the conference room without written direction: autonomous control without confirm. - We do not run AI in the conference room without written direction: AI summary to non-attendees. - We do not run AI in the conference room without written direction: retained recording absent policy. - We do not run AI in the conference room without written direction: outbound telemetry on any model. - We do not run AI in the conference room without written direction: internet-routable AI endpoint.	

§05	AUDIO CAPTURE HARDWARE
- Sennheiser TeamConnect Ceiling 2 (509161) — beamforming ceiling array · Dante · TruVoicelift. - Shure MXA920 — square steerable array · 8 lobes · Dante/AES67 · IntelliMix DSP onboard. - Shure MXA710 — linear ceiling/wall array · 2 ft or 4 ft · 8 lobes · Dante/AES67. - Biamp Parlé TCM-XA — Beamtracking ceiling mic with onboard amp · TesiraFORTÉ DSP host. - Mic feed terminates at room DSP first — DSP gates AI capture path on consent signal only. - Intercom audio (Vol 10) never feeds AI capture path unless added to consent list — DSP enforces.	

§06	VOICE CONTROL INTEGRATION
- Crestron 4-Series · SIMPL# Pro HttpClient · POST audio → Whisper → POST text → Ollama /api/chat. - Crestron alt: Python on 4-Series → Ollama direct → EISC signal back to SIMPL — newer firmware. - Lutron HomeWorks LIP · Telnet 23 · #OUTPUT,[ID],1,[Level],[Fade] · 100 ms inter-command floor. - Lutron new build: LEAP API — preferred over legacy LIP · dynamic config · secure transport. - Control4 DriverWorks SDK · Lua · .c4i driver · maps intent JSON to Light V2/Blind/Thermostat proxies. - Latency budget 2 s: STT 200-400 ms · LLM 500-800 ms · intent 50 ms · scene 100-200 ms · ack tone <500 ms.	

§07	DOCUMENT Q&A AND RAG
- BGE-M3 embedding · 1024 dim · multilingual · dense + sparse + multi-vector · 8K context. - nomic-embed-text alt · 768 dim · Apache-2.0 · ollama pull nomic-embed-text · faster ingest. - ChromaDB local-first store · persistent on appliance volume · no cloud sync · Apache-2.0. - Qdrant alt for higher-volume deployments · Rust core · self-hosted Docker · quantization support. - LlamaIndex ingestion: SimpleDirectoryReader → SentenceSplitter → embed → Chroma upsert; rerun on doc change. - Document ACL stored locally · access tied to Vol 09 credential · only authorized docs returned per session.	

§08	NETWORK ISOLATION
- Dedicated AI VLAN — no DHCP relay to client VLAN · no inter-VLAN routing except control bridge. - Firewall: WAN egress DENY · DNS internal only · NTP internal only · NDP/mDNS scoped to VLAN. - Control bridge ACL: only Crestron/Lutron/C4 processor IPs may POST to AI appliance ports. - DNS telemetry blocking: Pi-hole or pfBlockerNG list including model and platform vendor domains. - Internal TLS cert for Open WebUI · self-signed CA distributed to client devices · no Let's Encrypt. - Validation: ping from each user VLAN must fail · curl https://example.com from AI must time out.	

§09	PRIVACY & COMPLIANCE
- GDPR Ch 5 · Art 44-50: local-only inference avoids third-country transfer analysis entirely. - HIPAA Security Rule (45 CFR 164): local processing keeps PHI on premises; no BAA gap. - Attorney-client privilege: work product never leaves the room · no third-party processor exposure. - SEC 17 CFR 275, Rule 204-2: family office advisory communications retained per client policy, on-site. - NIST AI RMF · MAP + MEASURE: local model removes class of cloud data-handling risks. - C2PA v2.4: AI-generated meeting summaries optionally stamped with provenance manifest before export.	

§10	CAPTURE & CONSENT
- NJ N.J.S.A. 2A:156A-3: one-party consent · participant may record · signage at room entry. - CA Penal Code 632: all-party consent · verbal confirmation required before transcription begins. - Two-party states: CA, FL, IL, MD, MA, PA, WA — green-button consent at panel before capture. - Standard NJ disclosure script: "AI transcription may be active. Audio does not leave the building." - Standard CA script: "Transcription requires consent of all participants. Press green button to confirm." - Calendar invite footer: "Restrepo AI active in this room. Local only. Contact admin with questions."	

§11	CLOUD AI COMPARISON
- Otter.ai: cloud transcription · audio uploaded · stored · indexed · subpoena-reachable. Not approved. - Zoom AI Companion: cloud LLM · meeting content processed off-site · not approved for boardroom. - Microsoft 365 Copilot: enterprise data protection but cloud-resident · approved only with EDP + DPIA. - Apple Intelligence (Private Cloud Compute): off-device path exists · not approved for confidential rooms. - Google Gemini: cloud-resident · Workspace Enterprise controls · not approved for Vol 15 use case. - Restrepo Vol 15: local-only inference is the standard. Cloud AI is a different product, not a tier.	

§12	CROSS-VOLUME COORDINATION
- Vol 09 Access Control: credential at door yields identity token to AI appliance · sets RAG ACL scope. - Vol 10 Intercom: DSP gates intercom audio out of AI capture path absent consent list inclusion. - Vol 11 Cameras: "room occupied" starts session · "room vacant" flushes buffers · PIR redundant gate. - Vol 12 HT Installation: AI rack shares cabinet only when thermal and acoustic isolation engineered. - Vol 13 HT Calibration: room acoustic data informs Whisper noise profile · improves WER under HVAC. - Vol 14 AI in the Field: same hardware/software · field crew familiarity transfers to client install.	

§13	QC · AUDIT · CLOSEOUT
10-item QC gate — all 10 PASS before client handoff or operator sign-off. - Naming: ProjectName_Vol15_AIConfQC_YYYYMMDD.pdf · privacy posture sheet signed by client. - Spare hardware ≥5% delivered + labeled · spare Mac Mini for hot-swap on Tier 2 builds. - Owner walk-through: demonstrate consent flow, voice command, document Q&A, kill-switch on panel.	

QC GATE — VERIFY BEFORE CLOSE-IN

☐ VLAN isolated — ping from user VLAN to AI appliance fails	☐ No WAN egress — curl public URL times out at firewall	☐ Model SHA verified — sha256sum matches published digests	☐ Transcription tested — WER under 5% on known phrase	☐ Voice commands tested — 3 commands within 2 s budget
☐ RAG tested — test doc queried, response under 60 s	☐ Disclosure played — consent script triggers in occupied room	☐ Retention policy on file — signed by client	☐ Privacy posture sheet signed — client signature on file	☐ Closeout package — QC log, consent, digests, scripts filed