Reproducibility configuration

Every value extracted from the repository's code and configuration; file references are to the repository at the archived release (v1.0.0, DOI 10.5281/zenodo.22689261), whose lockfile (uv.lock) pins every software dependency to an exact version. Deterministic settings governed both the harness runs and the live room runs (the live environment sets no overriding values).

Table S1. Reproducibility configuration.

Item	Value	Source
Clinical LLM	MedGemma 27B, Q4_K_M GGUF (hf.co/unsloth/medgemma-27b-text-it-GGUF:Q4_K_M), one tag at all five call sites	app/cds.py; notes.py; rag.py; letters.py; finalize.py
Embedding model	embeddinggemma (768-d)	app/rag.py
Context window	16,384 tokens shared by CDS/RAG/notes (single value so the model never reloads); letters 8,192	app/cds.py; app/letters.py
Sampling	temperature 0.0, seed 42, everywhere (assessment, urgency, affect, notes, letters, RAG); env-overridable, unset in the live environment	app/cds.py and call sites
Other model options	keep_alive 30 m on chat calls; structured-output schema enforced; MedGemma evicted (keep_alive 0) before finalisation	per call site
Live ASR	faster-whisper distil-large-v3; beam 5; language en; VAD filter on; condition_on_previous_text off; clinical initial prompt	app/transcription.py
Final-pass ASR	WhisperX large-v3, float16 on CUDA, silero VAD, batch 8; WhisperX 3.8.6	app/finalize.py; uv.lock
Diarisation	pyannote/speaker-diarization-3.1; package 3.4.0	app/finalize.py; uv.lock
TTS	Piper, voice en_GB-alba-medium, CPU-only isolated subprocess	app/speech.py
Note grounding gate	refuse when cited fraction < 0.5, or zero claims	app/notes.py
Load-bearing flag	numeric/drug/dose/laterality claim with any cited turn's ASR confidence < 0.60 → flag	app/notes.py
Transcript-quality gate	S2 confidence: refuse < 0.60, flag 0.60–0.70; S3 repetition: flag ≥ 0.5 within-segment share (≥12-token segments); S4 trailing loss: refuse on measured ≥ 3.0 s untranscribed run (20 s	app/ transcript_quality.py

	fallback when unmeasurable), flag at 10 s; S1 language: measure-only	
Retrieval	similarity floor 0.45, top-k 4; refusal below floor; uncited answers rejected	app/rag.py
Audio	16 kHz mono 16-bit PCM WAV live; lossless FLAC on approval	app/live.py; app/retention.py
Host	RTX 4090 (24 GB); Ubuntu Linux, run under WSL2 on Windows 11 for the evaluation reported	manuscript, Methods
Initial development	213 commits on 12 active days over 27 days (2026-07-05 to 2026-07-31)	git history