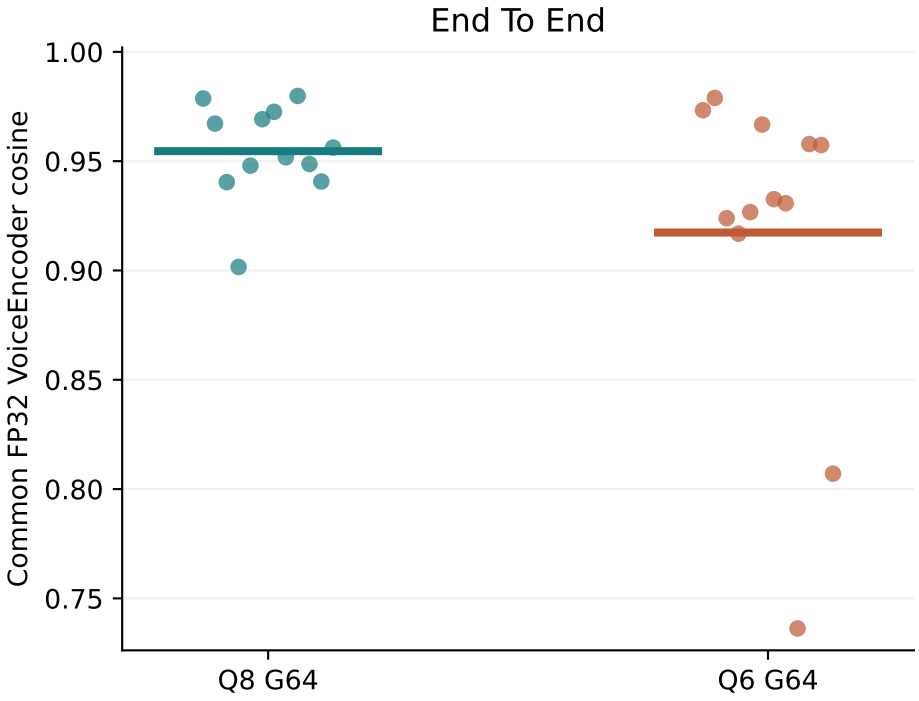
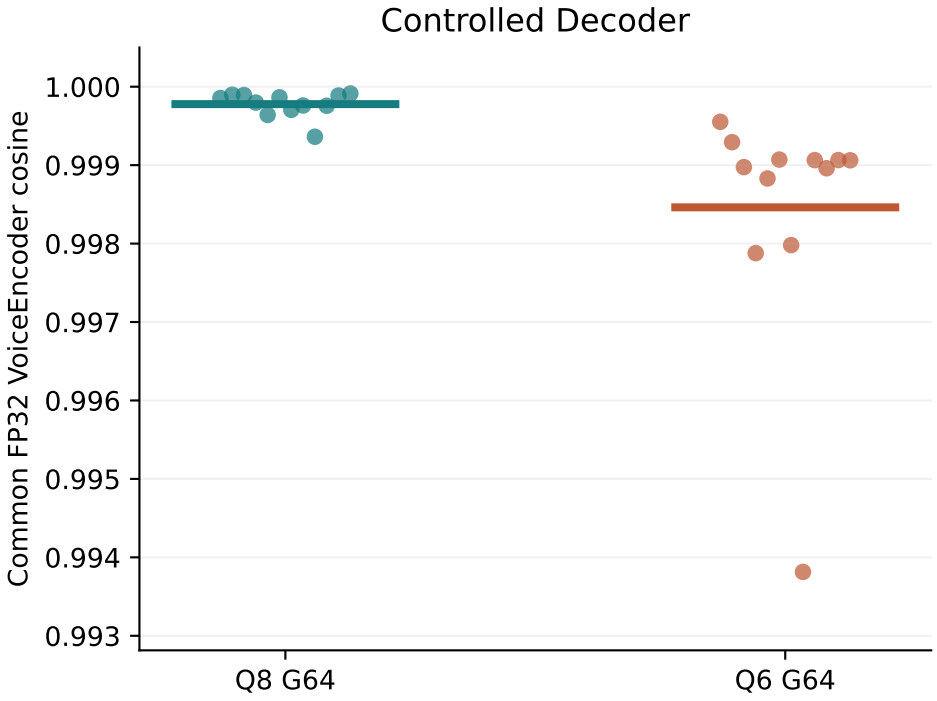
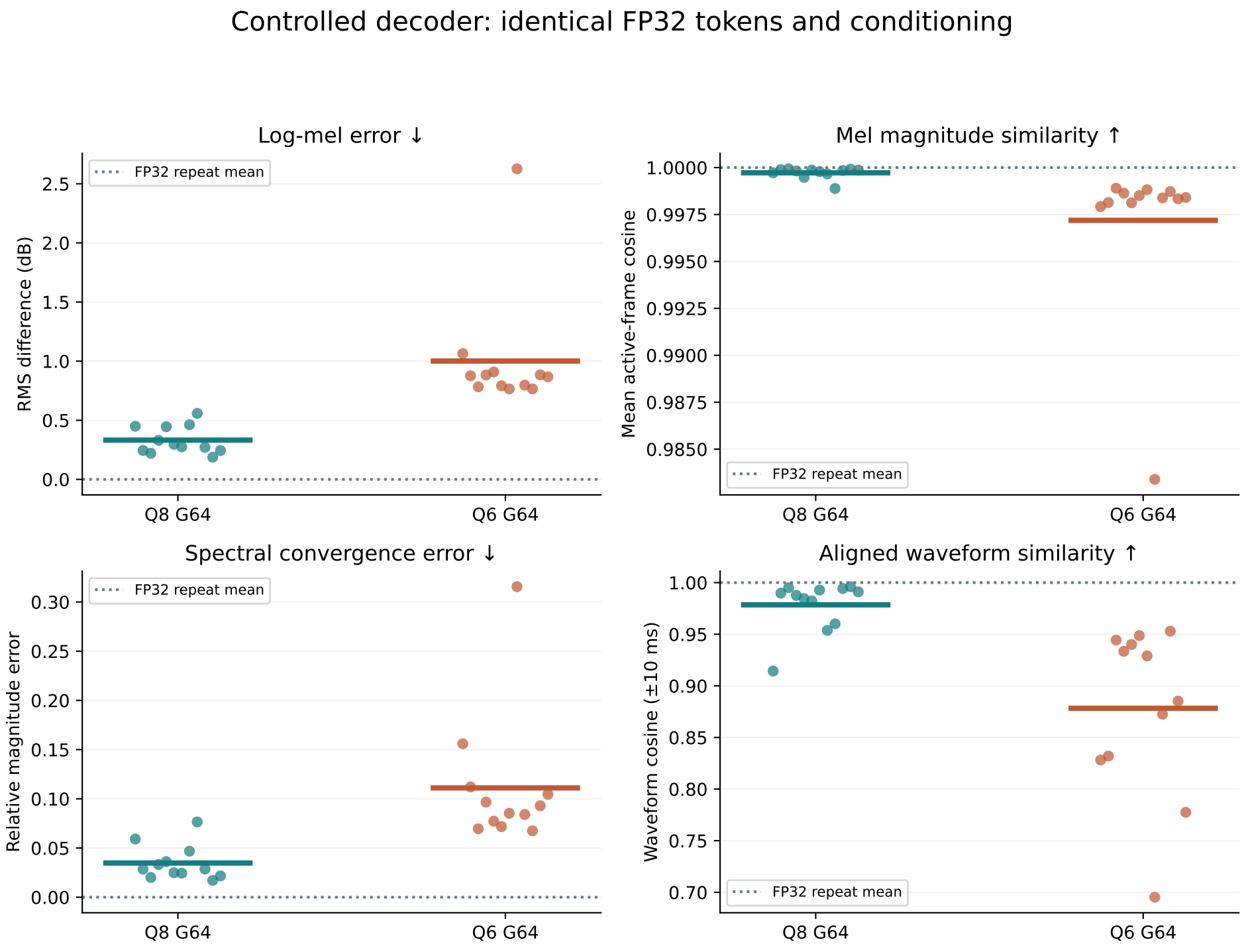
Voice identity versus matched FP32 output

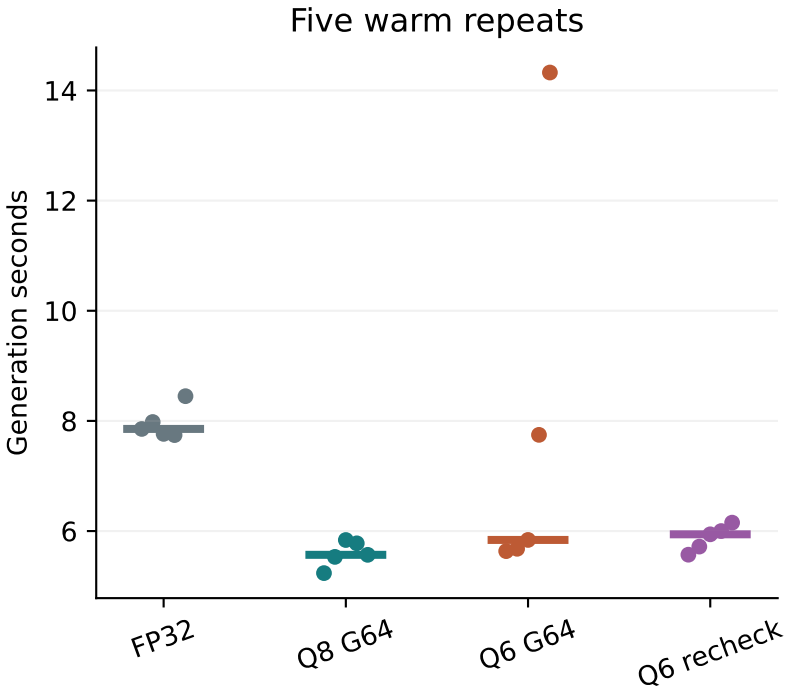


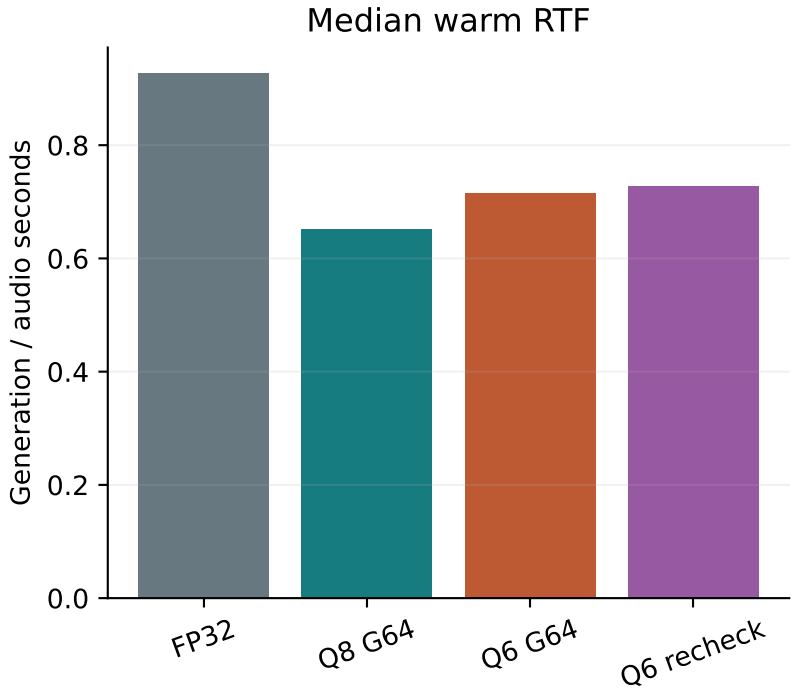


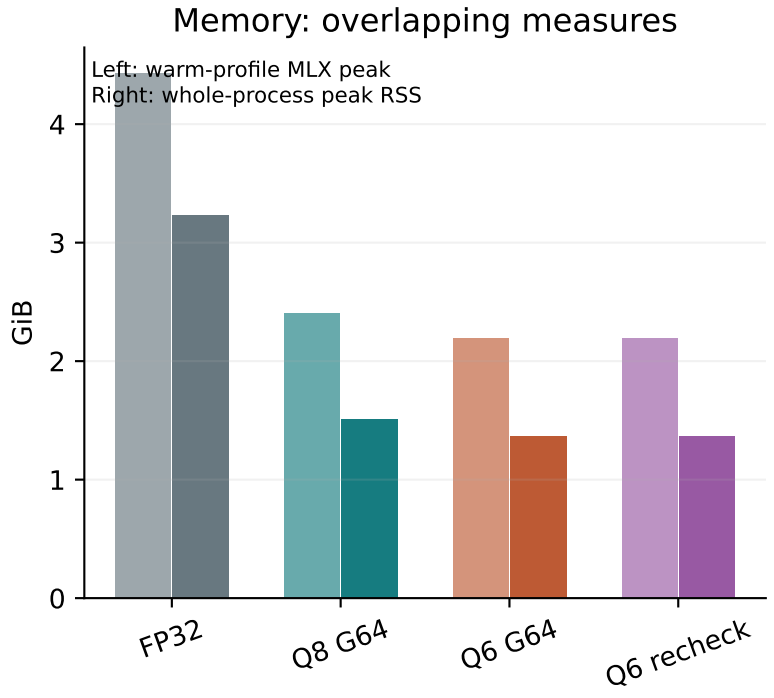
Dots: 12 clips per candidate; bars: means. Panel scales differ. In-family identity proxy, not MOS.


Controlled decoder: identical FP32 tokens and conditioning
Log-mel error ↓
RMS difference (dB)
FP32 repeat mean
Q8 G64
Q6 G64
Mel magnitude similarity ↑
Mean active-frame cosine
Spectral convergence error ↓
Relative magnitude error
Aligned waveform similarity ↑
Waveform cosine (±10 ms)
Dots: 12 clips; bars: means. Dotted line: 12 FP32 self-repeats. Waveform scores are sensitive to phase.

Local generation cost: measured in sequence on battery

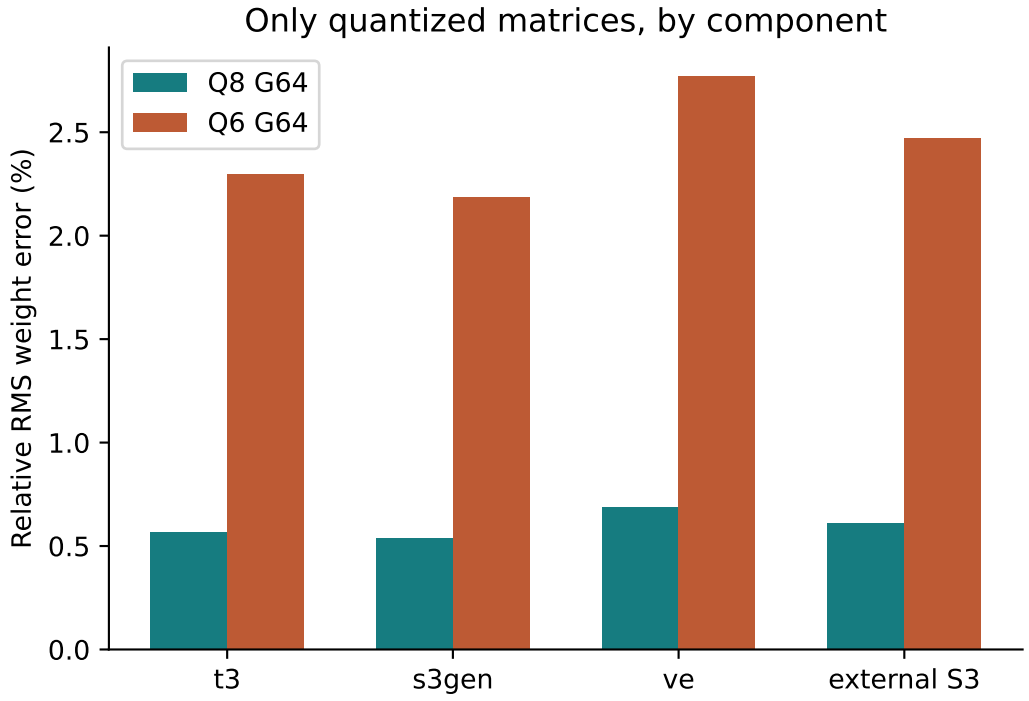


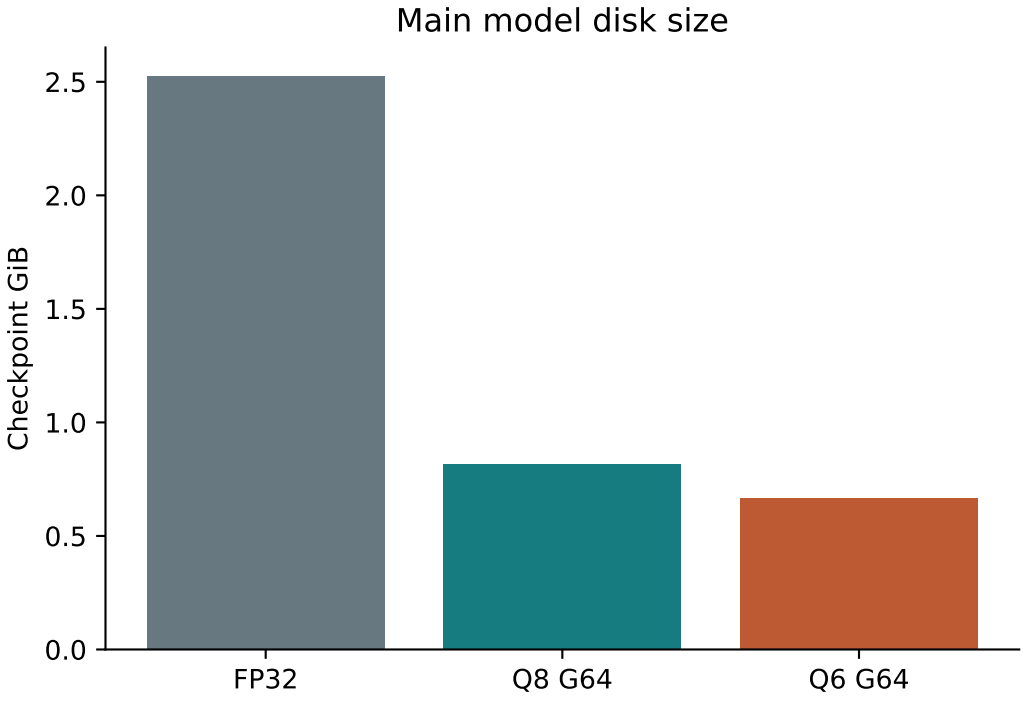




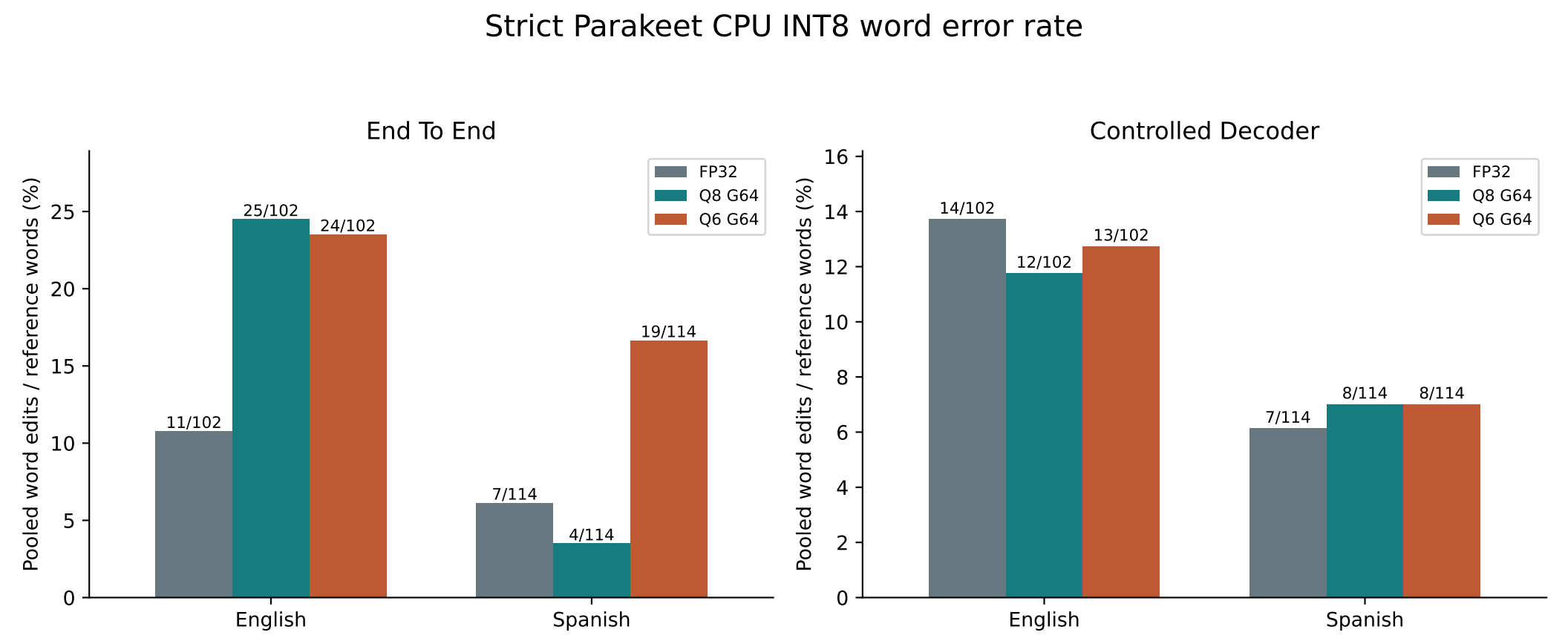
Bars in timing panel: medians. Live desktop/background activity affects timings; MLX and RSS must not be added.

Weight precision audit against the original FP32 checkpoint





651 main matrices + 37 runtime S3 matrices. All 1,836 other main tensors (including convolutions) are byte-identical.


Strict Parakeet CPU INT8 word error rate
End To End
Controlled Decoder
Pooled word edits / reference words (%)
FP32
Q8 G64
Q6 G64
11/102
25/102
24/102
7/114
4/114
19/114
14/102
12/102
13/102
7/114
8/114
8/114
English
Spanish
Three phrases per language × two seeds. Acronyms/numbers are not expanded. ASR errors are not all TTS errors.