electronics

MDPI

Article

Latency, Memory, and Verifiable Spoken-Format Constraints in Local ASR–LLM–TTS Pipelines: An Open-Domain Factorial Evaluation

Balázs Gáspár [1,2,†], Kata Egres [1,†], Ádám Pintér [1,†], Zsolt Bringye [1] and Gábor Kertész [1,2,*]

[1] John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary; gaspar.balazs@nik.uni-obuda.hu (B.G.); egres.kata@nik.uni-obuda.hu (K.E.); pinter.adam@nik.uni-obuda.hu (Á.P.); bringye.zsolt@nik.uni-obuda.hu (Z.B.)
[2] Laboratory of Parallel and Distributed Systems, Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 1116 Budapest, Hungary
* Correspondence: kertesz.gabor@nik.uni-obuda.hu
† These authors contributed equally to this work.

Abstract

A spoken reply requires a host-side automatic speech recognition (ASR), language model (LLM), and text-to-speech (TTS) pipeline. Latency, device memory, and spoken-format adherence are properties of that pipeline. We evaluate it on 999 English read-aloud HeySQuAD questions under two recognizer regimes. Each turn is question-only: the LLM receives the recognized question text and a fixed system prompt, without the SQuAD passage. Scoring is text-side on archived replies plus synthesis timing; no listening test or conversational usefulness rating is claimed. Conversational latency is speech-end-anchored time-to-first-audio (TTFA). Format pass is the strict conjunction of programmatic hard checks on the short-opener prompt; the five-word opener dominates failures. It is not a general instruction-following or task-achievement score. On a measured validation subset under streaming Vosk, TTFA medians were 1.28–1.54 s (2.03 s on a 27B cell); endpoint delay dominated the wait except on that cell. Absolute TTFA is not reported for the buffered Whisper quality arm. Corpus word error rates were 0.200 on streaming Vosk and 0.049 on Whisper. Four-bit quantization reduced device memory by a factor of about two to three. Under the full conjunction, Q4 raised format pass on Llama, showed no detected change on Qwen, and lowered it on Gemma, and majority pass occurred only at 27–32 B. Without the five-word opener check, both effects change: all twelve temperature-0 Vosk cells pass on 88.8–99.6% of items, and the Llama Q4 gain and Gemma Q4 loss are gone. Smaller models could satisfy the full conjunction while showing low coverage of the intended question across all items. Short-answer reference-span presence remained low under the question-only prompt. Under the higher-error regime, intended-question coverage fell more than format pass.

Keywords: large language models; on-host speech pipeline; local inference; model quantization; automatic speech recognition; text-to-speech synthesis; LLM evaluation; spoken-format constraints; time-to-first-audio

Check for updates

Academic Editor: Zihao Lei

Received: 19 September 2026
Revised: 19 October 2026
Accepted: 26 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Spoken interaction needs the automatic speech recognition (ASR), language model (LLM), and text-to-speech (TTS) stack that menus and text-only prompts avoid. Recognition

must run while the user is speaking, and synthesis can start on the first speakable sentence while tokens are still arriving [1]. The wait the user perceives is then dominated by endpointing, the first synthesizable text, and first-chunk synthesis.

On-host serving keeps speech and derived text under local control, consistent with data-protection-by-design principles [2], and keeps the experiment reproducible without third-party APIs. Model size, device preset (CPU or GPU), and quantization then matter directly: they shape both time-to-first-audio (TTFA) and the quality of the spoken reply.

Text-only agent studies score reply quality; systems papers often time ASR, TTS, or an LLM alone. The object here is the locally served cascade: engines and compression vary, the measurement protocol does not. Open-domain spoken question answering is the first layer for latency, device memory, throughput, and verifiable spoken-format constraints under recognizer noise.

This paper evaluates that cascade on a shared open-domain English read-aloud corpus. Each turn is question-only: no SQuAD passage is supplied to the model. We report speech-end-anchored TTFA after endpointer calibration, stage latencies, device memory and throughput, and text-side format pass under a hard-check conjunction, including behaviour under two recognizer regimes. The design uses two recognizer arms and three interleaved launches per cell, where a cell is one combination of model tag, decoding temperature, and recognizer arm under fixed synthesis. Scoring is text-side on archived replies plus synthesis timing.

The architecture was initially motivated by a possible future use in therapist-supervised digital training, but the present study evaluates a general-purpose local spoken assistant rather than a therapeutic system. The experiments use open-domain English read-aloud questions and do not involve therapists, clinicians, or patients.

1.1. Research Questions

RQ1: What speech-end-anchored TTFA is observed on a measured validation subset of locally served LLM configurations under a fixed ASR–TTS path, and how do stage components compare across the broader factorial?

RQ2: How do the tested model tags, sizes, and quantization settings associate with response time, generation throughput, and device memory (VRAM)?

RQ3: On question-only English read-aloud items, how often do the evaluated LLMs satisfy the spoken-format hard-check conjunction (opener-dominated; not general instruction following or usefulness) when scored text-side from archived replies?

RQ4: What trade-off exists between computational efficiency and that format-pass rate (with lexical surrogates reported separately, not as conversational quality)?

RQ5: Under two compound recognizer regimes on the same recordings (streaming higher-error Vosk versus buffered low-error Whisper), how do format pass and lexical contact with the intended question change?

1.2. Contributions

- A reproducible on-host ASR–LLM–TTS measurement protocol with speech-end–anchored TTFA, dual-reference Vosk endpointer calibration, split recognition/generation with reconstruction validation, and a screened open-domain read-aloud corpus (999 validated items). On a 200-item validation subset under streaming Vosk, measured TTFA medians were 1.28–1.54 s (2.03 s at 27 B).
- A factorial of locally served LLM configurations under fixed Piper synthesis (24 Vosk and 12 Whisper cells, each launched three times), reporting measured TTFA on five validation cells, reconstructed TTFA and stage latencies on the factorial, device memory

and throughput, and text-side format pass. Whisper is the frozen-transcript quality arm, not a TTFA condition (Section 3.3).

- An analysis that separates opener-dominated format pass from lexical coverage of the recognized versus intended question under two compound recognizer regimes (corpus WER 0.200 streaming Vosk; 0.049 Whisper), including a latency–format trade-off under that hard-check conjunction and a short-answer exact-span presence indicator interpreted under the question-only prompt.

2. Related Work

2.1. Systems Motivation: Local Spoken Assistants

Local spoken assistants can support hands-free interaction while keeping audio and derived text on the evaluation host. Recognition, generation, and synthesis jointly set measured latency and spoken-format adherence. This study uses open-domain speech to isolate those pipeline properties. The closest cascade evaluations often report recognition WER, component latency, or end-to-end wall time. Fewer also fix a speech-end anchor, calibrated endpointing, first-audio TTFA, reconstruction checks for a split design, dual recognizer regimes, and a factorial of local LLM tags under one synthesis path. We contribute that measurement package and the factorial under it. We do not introduce a new recognizer or generative architecture. Claims beyond this host and these settings stop at the measurement boundaries in Section 3.

2.2. Local Speech Pipelines and Conversational Latency

Interactive speech systems need recognition accuracy and a wait that remains usable in conversation. Multi-second delays degrade experience in controlled studies, although no universal voice-interface standard fixes a single second-scale target [3–5]. ASR is commonly scored with word error rate (WER). Interactive use also needs endpointing, because emission and end-of-speech behaviour enter user-perceived latency, and real-time factor or FLOPS need not predict deployed wait [6–8].

Whisper is a general-purpose recognizer [9], but its original design targets complete segments rather than continuous low-latency interaction [10]. Offline engines such as Vosk support streaming recognition with compact models [11]; buffering and endpoint detection must be set carefully. On the reply side, neural TTS systems are often scored with listening tests or mean opinion score on controlled samples [12–15]. Piper provides lightweight offline synthesis [16]. In a speech-to-speech pipeline, first-chunk synthesis latency enters the perceived wait directly. This study times synthesis; it does not report perceptual TTS quality.

2.3. Local and Compressed Large Language Models

Local inference keeps model weights and audio on the evaluation host, which reduces dependence on external services and keeps data under local control, consistent with privacy-by-design principles for sensitive data processing [2]. Memory and compute demands still limit which models fit. Surveys of inference efficiency cover reductions in model size, memory, and compute and generation time [17]. Frameworks such as Ollama and llama.cpp serve open models on CPU and GPU [18,19].

Quantization lowers weight precision to cut memory. Common bit-widths include 8-, 6-, 5-, and 4-bit representations. Surveys of low-bit LLMs report large memory and compute savings that can allow larger models on constrained hardware [20]. In llama.cpp, the `GGUF` format and several quantization schemes implement these widths [19]. Gains depend on the model, the hardware, and the quantization method, so the lowest bit-width is not always the best choice.

Lower precision does not guarantee faster generation. Jin et al. report that 4-bit models can match higher-precision accuracy on many tasks while cutting memory, and that inference can also slow under some settings [21]. Comparisons of local LLMs therefore need both runtime and response measures, not size or bit-width alone.

2.4. Instruction-Following Evaluation Under ASR Noise

Verifiable instruction-following benchmarks score whether outputs satisfy stated constraints [22,23]. Recognizer choice changes architecture, decoding, timing, and error type as well as WER; WER alone does not determine semantic task performance [24–26]. This evaluation separates lexical coverage from a hard-check format-pass conjunction under two such regimes.

SpeechGPT [27], Moshi [28], LLaMA-Omni [29], and Mini-Omni [30] optimise a different boundary from a controlled on-host ASR–LLM–TTS cascade with speech-end–anchored TTFA. Local and quantized models are mainly compared on memory and throughput [17,20]. Spoken reading-comprehension and dialogue corpora under ASR noise quantify content loss [24–26]. They do not fix the first-audio measurement and reconstruction checks used here. We measure the modular cascade under one protocol: format pass scored separately from contact with the intended question, under two recognizer regimes, with TTFA anchored at speech end.

3. Methodology

The experimental object is a modular, locally served ASR–LLM–TTS pipeline. This section specifies its execution modes, corpus construction, measurement boundaries, configuration design, and repeated-launch protocol. Low-level settings and preparatory calibrations are collected in Appendix B; response scoring is specified in Section 3.10.

The design uses 1001 eligible recordings, recognizer-specific frozen transcripts, and independent pipeline launches. It separates model-size and decoding comparisons from within-model weight-precision comparisons. Timing and text-side response measures are defined independently so that recognition fidelity, response-format compliance, and computational cost are not treated as interchangeable outcomes.

3.1. System Architecture

Recognition, language generation, and synthesis execute as cooperating workers in one Python 3.12 pipeline process (Figure 1), with the language model served by a separate Ollama process on the same host. Audio and text remain local; the language-model client communicates over loopback HTTP. Each recording is an independent request under the same system prompt. Neither previous questions nor previous replies are retained as dialogue history. The corpus context and reference answer are evaluation fields, not additional information supplied to the model.

The full path is concurrent rather than a sum of three batch operations. The ASR worker consumes audio chunks and publishes finalized text. The LLM worker sends that text to the local server and receives a token stream. As soon as a speakable chunk is complete, it is placed on the TTS queue; synthesis can therefore overlap the generation of later tokens. The first synthesized chunk is the first-audio event. Saving or playing that waveform occurs after this boundary.

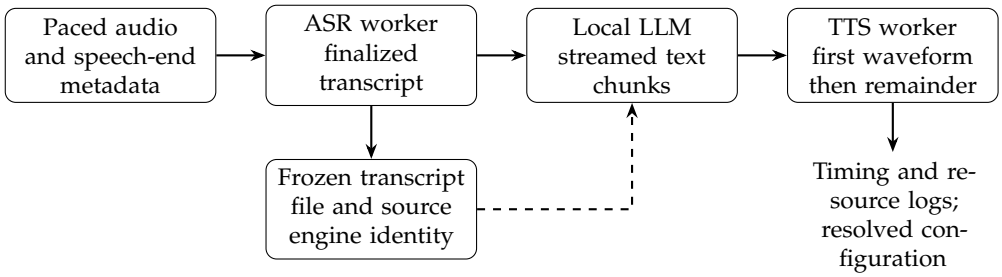


Figure 1. Implemented execution paths. Full runs use the upper path. ASR-only runs stop at the transcript archive; text-input runs enter at the LLM using that archive. Dashed routing denotes a separate invocation, not simultaneous feedback. The timing log records worker-specific boundaries rather than assigning all elapsed time to one sequential stage.

3.1.1. Full, ASR-Only, and Text-Input Execution

The three modes expose the same components through different boundaries (Table 1). Full execution is the reference for directly measured TTFA. ASR-only runs archive finalized transcripts and, where the time base permits, endpoint delay. Text-input runs consume a frozen transcript file and record downstream latency and replies, but cannot measure speech-end–anchored TTFA directly.

Table 1. Execution modes and their measurement boundaries. A blank quantity is unavailable, not a zero-duration stage.

Mode	Input	Active Components	Latency Available
Full	Audio; optional timing metadata	ASR, LLM, TTS	Component stages and direct TTFA when anchored
ASR-only	Audio; optional timing metadata	ASR	Recognition and anchored endpoint delay
Text-input	Frozen transcript file	LLM, TTS	Generation and synthesis; reconstructed TTFA only after joining

Split recognition keeps every LLM cell in a transcript arm on identical wording and allows a second recognizer without re-decoding every cell. Split execution is not physically equivalent to concurrent full runs: contention and worker scheduling differ, so reconstructed TTFA remains an estimate requiring the validation in Section 3.9. Transcript records identify the source file, recognized text, and engine when present; the exact frozen file must be retained because a path is not a content checksum.

3.1.2. Endpointing, Superseded Answers, and Failure Handling

Only non-empty finalized recognition results trigger generation. A later finalized continuation cancels the answer to a shorter provisional input; the retained TTFA uses the surviving request and its first synthesized chunk. Playback is disabled in timed file-input configurations. Failed LLM requests yield a blank reply with an error marker rather than fabricated stage success; empty recognition is recorded as skipped. Missing stage rows and incomplete items remain visible in analysis denominators.

3.2. Hardware and Software Environment

The evaluation environment is a single virtual machine on HUN-REN Cloud [31], a Hungarian science cloud infrastructure, with the specifications listed in Table 2. The pipeline and the Ollama server share this host. A pipeline device preset does not by itself determine the location of the LLM weights: recognizer and synthesizer settings are resolved in the client, whereas Ollama manages language-model execution independently.

Table 2. Evaluation-host specification and fixed language-model serving environment.

Attribute	Specification
CPU allocation	16 vCPU
System memory	64 GB
GPU	NVIDIA GRID V100DX-32C, 32 GB
Compute capability	7.0
NVIDIA driver	580.65
Operating system	Ubuntu 24.04.3 LTS
Python	3.12.3
Ollama	0.33.3

Ollama 0.33.3 is held fixed across configuration comparisons. Client-observed pre-token latency includes prompt evaluation and serving overhead, which are not solely properties of the model weights. Server version, prompt framing, cache state, and residency are controlled or recorded alongside model settings.

For reproducibility, each launch retains `config_used.yaml`, its input provenance, stage log, and output artefacts (Appendix B.6). The evaluation record additionally identifies the server version, model artefacts, and environment used for those launches; a mutable model tag or a requirements file alone is insufficient.

Models are initialized before timed items. The LLM warmup uses the same system-prompt framing and runtime options as measurement, and the server is asked to retain the model in memory. The server was configured to hold one model at a time (`OLLAMA_MAX_LOADED_MODELS=1`), so a newly loaded model never shared the accelerator with its predecessor. In all 159 language-model launches the server reported the model as 100% GPU-resident, and device memory in use exceeded that model's allocation by the same amount in every launch of a cell, although a different cell preceded it in each round. Warmup and residency checks reduce cold-start contamination but do not prove that every subsequent request has identical cache state. Residency diagnostics must be read with the run outputs. The first input of each independently launched run is excluded from warm latency summaries to limit residual lazy-initialization effects; it remains in the raw artefacts. These are warm pipeline measurements, not cold-start or server-restart benchmarks.

3.3. Speech Recognition Component

Audio is normalized to mono 16 kHz, 16-bit PCM. The audio source yields fixed-duration chunks to either recognizer. Under real-time file pacing, delivery follows the audio time axis; under fast pacing, chunks are supplied without waiting. Only the former can map a speech-end offset within a file to a wall-clock latency. This is a controlled replay interface, not a claim that read-aloud recordings reproduce spontaneous conversation.

Vosk supplies the streaming arm, using `vosk-model-small-en-us-0.15` on CPU [11,32]. It produces partial and final hypotheses and exposes word timestamps. A fresh decoder is prepared before each timed stream; reusing a decoder's accumulated sample clock would invalidate file-relative anchors. The selected chunk size is 100 ms and the flat endpoint-silence setting is 600 ms. These are recognizer settings, not an LLM decision about whether the user has completed a sentence. For the reported evaluation they were held fixed,

without a Vosk model upgrade or endpointer retuning. Their preparatory calibration and its limits are given in Appendix B.

The comparison arm uses faster-whisper, a CTranslate2 implementation of Whisper [9,33,34]. In this pipeline it buffers the entire audio stream and decodes after stream completion. This buffered path is the low-error transcript-quality contrast for RQ5, not a second streaming-TTFA condition; file duration and trailing silence enter its release time differently from Vosk. The evaluated configurations specify model `small`, CUDA execution, and an explicit `int8` compute type, which overrides any generic device-dependent default (including CUDA `float16`). The `small.en` model used in the separate corpus-mismatch screen is not the pipeline's `small` configuration.

Both recognizers are evaluated on the same eligible recordings and reference questions. WER characterizes recognition fidelity within each regime, while downstream frozen-transcript arms characterize the consequences of that regime's transcript text. The Vosk–Whisper contrast is a compound recognizer-regime contrast (architecture, decoding, timing, and error type), not a pure error-propagation experiment that holds all other factors fixed. No reference-question (oracle-text) arm was run, nor does lower WER alone guarantee better answers or better format adherence.

3.4. Locally Served LLMs

The client streams responses from Ollama [18]; locally stored model variants use the serving stack's supported quantized formats [19]. The model-size and decoding comparison comprises 24 Vosk configurations and 12 Whisper configurations, with three independent launches per configuration (Section 3.11). A weight-precision comparison covers nine model–precision combinations at temperature 0. This configuration space is not a balanced cross-product of every recognizer, precision, and decoding setting. Each comparison uses an explicitly selected cell set and recognizer-derived input arm. Table 3 gives the model inventory, with exact tag construction and the configuration matrix in Appendix B.

Table 3. Language-model families and weight precisions in the configuration space. Precision contrasts hold model identity and parameter count fixed; the three architecture families are not interchangeable replicates.

Family	Parameter Counts at Q4_K_M	Additional Precisions
Llama	1 B and 3 B (3.2); 8 B (3.1)	8 B: Q8_0 and FP16
Qwen 2.5	1.5 B, 3 B, 7 B, 32 B	7 B: Q8_0 and FP16
Gemma 3	1 B, 4 B, 27 B	4 B: Q8_0 and FP16

The weight-precision comparison uses Llama 3.1 8 B, Qwen 2.5 7 B, and Gemma 3 4 B, each in `Q4_K_M`, `Q8_0`, and FP16 form, with LLM temperature fixed at 0. Thus, nine denotes three models at three precisions, not the number of independent launches. FP16 is the higher-precision reference; the two quantized variants are compared with it within each model. Recognizer-derived input, system prompt, context and output-token limits, serving version, and synthesis settings are matched within each contrast.

Each analysis input identifies one designated set of run artefacts for the corresponding model–precision configuration. The same recording identities and transcript source define the paired items across precisions; additional copies or alternative runs of a configuration are not silently pooled into that comparison. Differences across parameter counts or model families remain model-configuration comparisons, not isolated estimates of parameter count. Similarly, a recognizer contrast changes both the text supplied to the LLM and the timing of that text's availability; those effects are separated through the transcript arms and stage decomposition.

All evaluated configurations cap generation at 150 tokens with a context setting of 1024. Greedy cells use a temperature of 0; sampled cells use 0.7 with seed 42. The matrix includes sampled Vosk cells but does not provide a sampled counterpart for every precision or recognizer cell. These controls constrain decoding but do not guarantee byte-identical outputs across separate launches when host or backend state can change. Such differences arise when the same request is processed in batches of different sizes, because floating-point reductions are not associative and their order in GPU kernels depends on the batch size and the hardware configuration [35,36]; here every request ran alone in the batch, in the same order, with unchanged server software, model placement, and GPU (Section 3.2), and sampled cells passed seed 42 with every request, so identical logits yield identical draws. Under the conditions used here (identical frozen transcripts, the fixed decoding settings above, warm Ollama server shared across launches, and interleaved client relaunches without server restart), the reported factorials produced byte-identical replies, and hence identical aggregate quality metrics, across the three launches of each cell (Section 4.10).

Text is released to synthesis at sentence-like punctuation boundaries, including full stops, question marks, exclamation marks, colons, semicolons, and newlines. A maximum chunk length of 140 characters bounds the waiting time for punctuation by allowing a split at a word boundary. Decimal-point handling prevents a simple numeric decimal from being mistaken for a sentence boundary; this is not a general abbreviation-aware linguistic segmenter. The first chunk depends on both the model's output and the segmenter's rules, so time-to-first-token and time-to-first-chunk are distinct quantities.

3.5. Text-to-Speech Component

The matrix holds synthesis constant with Piper [16] on CPU and 12 intra-operation threads. Voice and runtime options are resolved from each configuration. Synthesis is blocking within the TTS worker but concurrent with subsequent token generation. The first-chunk timer ends when that call returns a waveform, before waveform serialization or playback. Later chunks contribute to total synthesis time and total output duration but not directly to the first-audio boundary.

All evaluated cells keep Piper on CPU. Preparatory thread-count comparisons are summarized in Appendix B; this calibration motivates the fixed setting and is separate from the corpus-level model comparisons.

The timing instrumentation measures waveform production, not acoustic onset at a loudspeaker. Playback is disabled, and no listening test, intelligibility study, or perceptual speech-quality score is inferred from synthesis latency. Every cell uses the same Piper voice, settings, and CPU path. A listening test would rate that voice, not the LLM configurations; the cells differ in the synthesized text, which the text-side instruments score. A complete user-facing latency would additionally require buffering, device, and acoustic measurements beyond the defined boundary.

3.6. System Prompt and Conversational Boundaries

The common short-opener prompt is reproduced in Appendix A. It requests a concise, factual, English reply suitable for being read aloud: a first sentence of at most five words, followed only when necessary by one or two short sentences of detail, with no more than three sentences in total. It discourages formatting that is unsuitable for speech, asks the model to interpret obvious recognition errors silently, and permits a brief clarification when the request is unintelligible or critically ambiguous.

This instruction standardizes the intended response format and encourages an early synthesizable opener. An answer can satisfy a length constraint while missing the intended question, and a correct short answer may require no additional detail. No dialogue history

and no SQuAD passage are supplied to the model: generation is question-only from the recognized utterance plus the fixed system prompt. Context and gold-answer fields remain evaluation metadata for lexical checks only (Section 3.7). The prompt is held fixed across comparable cells, and its exact text must accompany the archived configuration.

3.7. Evaluation Corpus

The corpus is derived from the 4158-row HeySQuAD validation package [37], which contains spoken versions of questions in the SQuAD setting [38]. This number describes the source split used here, not the size of the complete HeySQuAD release. Selection consists of answerability filtering followed by audio–reference quality control. No pause-duration cutoff or random selection of easy-to-endpoint questions is applied.

First, the exact condition `is_impossible=False` retains all 1002 answerable records and excludes 3156 records marked unanswerable. The corresponding preparation report records preservation of row order and values, unique identifiers, agreement between CSV and ZIP order, archive-integrity checks, and byte-for-byte identity of the retained source WAV files. This verification concerns the answerability-filter output, before subsequent timing preparation.

Second, a conservative phonetic screen using an independent `small.en` transcription excludes one gross content mismatch, leaving 1001 recordings for the evaluation runs. During the runs, two further files failed waveform decode as truncated or unreadable WAVs (`572a0bfaaf94a219006aa77a.wav`, `5729e500af94a219006aa6b5.wav`). Exclusion rested on that decode failure alone; latency and response scores were not consulted. The two files may still appear in some archived run configurations. They are excluded from the 999-item scored denominators in the results (Table 4).

Table 4. Corpus-validity and analysis denominators. Answerability follows the source `is_impossible` flag; scored results use the 999 validated recordings. Exact-span presence uses the 982-item short-gold subset. The 1000 warm-item count is an upper bound before run-specific missingness when one of the 1001 eligible recordings is the first, excluded item. No SQuAD passage is supplied to the LLM (question-only generation).

Stage	*n*
Supplied HeySQuAD validation split	4158
Excluded by `is_impossible=True`	3156
Retained answerable questions	1002
Excluded by pre-run mismatch screening	1
Recordings used in evaluation runs	1001
Defective recordings identified during testing	2
Maximum warm timed items per launch without an additional warmup item	1000
Validated recordings (scored denominators)	999
of which short-span presence set ($\leq$8 gold words)	982
SQuAD passage in LLM prompt	0 (omitted)

Each record retains a filename and identifier, the intended question, source transcription, context and answer fields where present, and timing annotations. The intended question is the reference for recognition fidelity and intended-question coverage; the actual recognized text is the LLM input and the reference for recognized-input coverage. These roles are not interchangeable. Context and answer fields support text-side reference checks (including exact-span presence) but are never injected into generation. An answerability label inherited from reading comprehension means that an answer is available in the source context; it does not establish that an assistant given only the spoken question has enough information to reproduce that answer. Low gold-span presence under this question-only

protocol is expected for items whose recoverable answer depends on the omitted passage. Presence is not a passage-grounded reading-comprehension score.

The eligible metadata contains no missing intended questions or answer fields. All entries have a positive speech-end offset smaller than file duration, and the reported trailing silence agrees with duration minus speech end to within 1 ms. Median file duration is 7769 ms, median detected speech duration 4160 ms, and median trailing silence 2013 ms. The minimum recorded trailing silence is 1446 ms; the largest recorded internal pause is 1248 ms. These are metadata summaries, not a fresh waveform-level verification. They neither prove that every detector boundary is correct nor establish that every item survives every recognizer or launch. The two defective recordings identified during evaluation runs illustrate the limits of the preceding automatic screening.

The selection procedure avoids selecting items by the measured downstream latency or prompt-adherence outcome. It nevertheless remains a restricted corpus: English read-aloud factual questions, without spontaneous turn-taking, background interaction, or conversation history. Removing clear audio–reference mismatches protects measurement validity; it does not turn the remaining questions into a representative sample of real conversations. The corpus size follows complete answerable-set retention and the documented exclusions rather than a prospective power calculation.

3.8. Runtime and Resource Measurement

3.8.1. Speech-End Anchor and Real-Time Pacing

The start of perceived response latency is the end of the input speech, not the end of the WAV file or the time at which recognition is finalized. Let t_0 denote the stream's wall-clock origin and e_i the item-specific speech-end offset. Under real-time pacing the anchor is $t_{s,i} = t_0 + e_i$. The reported evaluation uses `speech_end_ms` from the accompanying metadata, derived from speech-region annotation rather than the recognizer being compared. A shared metadata anchor provides a better-controlled origin for engine contrasts than an engine-derived final-word boundary, although it remains an estimated speech boundary. Files with invalid or absent anchors require an explicit exclusion or separate analysis; an absent anchor is not replaced with zero.

Trailing silence allows Vosk to finalize before file exhaustion. It is not subtracted from recognition duration indiscriminately. Appendix B specifies the preparatory silence rule and its empirical origin. Fast-pacing runs are useful for processing-time measurements but do not supply speech-end–anchored wall-clock TTFA. A duration measured on fast replay cannot be converted to a real-time endpoint delay merely by subtracting the audio length.

3.8.2. First-Audio and Component Boundaries

For a surviving answer, let t_f be the final-text publication that supplies its input, t_q the downstream request start, t_c the release of its first synthesizable text chunk, t_b the beginning of that chunk's blocking synthesis call, and t_a the return of the resulting waveform. Direct TTFA is

$$\mathrm{TTFA} = t_a - t_s. \tag{1}$$

The observable critical path can be written as

$$\mathrm{TTFA} = \underbrace{(t_f - t_s)}_{\text{endpoint/release delay}} + \underbrace{(t_c - t_q)}_{\text{LLM time to first chunk}} + \underbrace{(t_a - t_b)}_{\text{first-chunk synthesis}} + \underbrace{(t_q - t_f) + (t_b - t_c)}_{\text{handover and scheduling}}. \tag{2}$$

This equation defines the boundaries; it does not assume that handover is identically zero or constant. The timestamp associated with the last non-empty finalized input need not coincide with the end of the recognition worker, which can continue draining the stream.

The log's `stt` duration covers that worker's timed stream-processing interval, including waits for audio under real-time pacing. It is consequently not the ASR contribution to TTFA. Similarly, complete LLM generation and total TTS synthesis overlap and extend beyond first audio (Table 5). Adding full recognition, full generation, and full synthesis times would double-count concurrency and measure a different quantity.

Table 5. Logged latency stages and their interpretation. Times refer to the surviving answer unless an event is explicitly marked as skipped or failed.

Stage	Boundary or Quantity
`stt`	Recognition worker's stream-processing interval
`stt_endpoint_delay`	Speech end to the relevant non-empty finalized input
`llm_ttft`	Client request to first generated token
`llm_prompt_eval`	Server-reported prompt evaluation
`llm_first_chunk_fill`	First token to first released text chunk
`llm_ttfc`	Client request to first released text chunk
`tts_first_chunk`	First chunk's blocking synthesis call
`ttfa`	Speech end to first synthesized waveform
`llm_eval`	Server-reported generation of all output tokens
`tts_total`	Sum of retained chunk-synthesis durations
`e2e_response_ready`	Item start to completion of the response-processing path

Time-to-first-token sits inside time-to-first-chunk rather than adding to it; server-reported prompt evaluation likewise sits inside the client pre-token interval. Generation throughput and recognition real-time factor are complementary diagnostics, not conversational wait. The end-to-end response-ready span includes input delivery and full-reply completion and is reported separately from TTFA.

3.8.3. Resource Observations

For RQ2 the reported resource outcomes are launch-level generation throughput and runner-reported LLM device allocation (VRAM). CPU utilization, process RSS, and GPU-utilization samples are logged beside stage windows; field definitions and reuse rules appear in Appendix B.5. Missing observations remain blank. Recognition, generation, and synthesis share the host, so those samples cannot be assigned to one stage and are not used as cell-level contrasts in results. They are host-load diagnostics only.

3.9. Reconstructed TTFA and Its Validation

Under frozen-transcript execution the speech-end–anchored wait is not observed directly. We reconstruct an item-level estimate by adding the canonical ASR endpoint delay to the downstream first-chunk path. For item i in downstream launch r,

$$\widehat{\mathrm{TTFA}}_{i,r} = E_i^{(a)} + C_{i,r}^{(x)} + S_{i,r}^{(x)} + h, \tag{3}$$

where $E_i^{(a)}$ is the endpoint/release delay from the selected ASR-only run, $C_{i,r}^{(x)}$ is downstream time-to-first-chunk, $S_{i,r}^{(x)}$ is first-chunk synthesis time, and h is an explicitly recorded handover correction. The endpoint term, not the full `stt` interval, carries the speech-side contribution. Each estimate is formed item by item before computing summary quantiles. A sum of separately computed medians is not a median reconstructed TTFA.

The reconstruction script exposes h as a configurable handover offset, with a default of 5 ms. This software setting is not a 5 ms standard deviation or an assumed bound on reconstruction error. The within-run residual check supplies that calibration quantity: $r_i = \mathrm{TTFA}_i - (E_i + C_i + S_i)$ diagnoses unassigned critical-path time, and the reported de-

fault is the median residual over the archived full-pipeline items used for the offset. A small residual supports boundary accounting but does not prove that removing a concurrent recognizer leaves the downstream terms unchanged. Waveform count alone is not a complete detector of premature endpoints. The split-versus-full experiment below tests transfer to separate execution. The offset can be set to zero to report the uncorrected sum; its chosen value and calibration basis accompany every reconstruction. Join rules, canonical ASR selection, and provenance checks used by the reconstruction helpers are summarized in Appendix B.6.

The primary transfer check compares reconstructed with directly measured TTFA on a common subset. A simple random sample of 200 recordings was drawn with seed 42 from the 1000 recordings that follow a warm-up recording, without forced inclusion of the hardest item and without stratification. `subset.json` in the validation archive records the draw and compares it with the corpus on speech-end–anchored endpoint delay, file duration, and trailing silence. The draw preceded the identification of the two defective recordings (Section 3.7); one of them was drawn and is excluded from all comparisons, leaving 199 items per cell. The measured arm ran the full pipeline under real-time Vosk pacing for five cells, three interleaved launches each, with the warm-up recording prepended to every launch. The split arm was not re-run on the subset: the reconstruction uses the factorial's canonical ASR-only pass and its three text-mode launches per cell, restricted to the subset items, so the check evaluates the reconstruction exactly as it is applied to the factorial. The two arms ran on different days, so host drift cannot be separated from split-versus-full differences. Absolute TTFA claims in Results refer only to this five-cell measured arm; the 999-item factorial contributes stage components and reconstructed TTFA.

For cells with repeated launches, the comparison uses item-level medians across launches within each arm before forming paired differences. Agreement is described through signed bias and the spread of those paired differences, alongside missingness and launch counts. The helper does not by itself establish statistical equivalence or remove uncertainty from between-launch variation. Reconstructed TTFA is reported separately from direct TTFA, and its agreement with full execution is assessed from this validation rather than assumed from the decomposition.

3.10. Spoken-Format and Lexical Scoring

RQ3–RQ5 are scored after the pipeline run, from the archived transcripts, by a separate evaluation package that does not import the speech workers. The unit is one (`stt_text`, `llm_text`) pair. Inputs are the open-domain items of Section 3.7; reference answers, when present, are used only for lexical overlap. Runtime and resource metrics are defined in Section 3.8. No human or domain-expert raters contributed to the reported scores, so no quality composite, safety gate, or acceptance verdict is claimed.

Each reported metric carries a validation status, so that a decidable check is not treated as a calibrated quality instrument:

- verifiable: decided by a deterministic programmatic check (without implying a flawless instrument or a pass rate free of sampling uncertainty);
- validated: a published instrument or sampling theory;
- surrogate: a deliberate simplification of a published method.

3.10.1. Verifiable Constraints and Readability

The constraint checks follow the verifiable-instruction design of IFEval [22]: each system-prompt instruction that can be decided programmatically is a named check. Item-level pass is the conjunction of all applicable hard checks; a check-level rate follows FollowBench [23] and omits inapplicable checks from the denominator. Reported results use the

strict item-level conjunction unless a caption states otherwise. Soft checks (spoken-duration budget, brief-enough length, and the conditional follow-up question) are advisory and do not count as hard prompt violations; markup bans remain hard checks.

The default constraint file is aligned with the short-opener English prompt for the hard checks that the program can decide: non-empty reply, opening sentence of at most five words ending in a full stop, at most two detail sentences after the opener (zero allowed, matching the prompt's conditional detail), English, no simulated dialogue, complete sentence-final punctuation, and no truncation at the recorded token cap. The prompt's conditional clarifying question is scored only as a soft advisory check, because critical ambiguity is not decidable from text alone; a question mark is therefore not counted as a hard prompt violation. Hard failures are the primary adherence outcome for RQ3. Item-level format pass is that decidable hard-check conjunction, dominated by the five-word opener. Its estimand, for each configuration, is the probability that a reply meets all hard checks; the observed pass rate over the scored items estimates it. The opener check matters for latency: the opener is usually the first text chunk sent to synthesis, so a shorter opener is synthesized faster and the first audio is ready sooner (Section 4.2). RQ3 asks how often each configuration meets the hard checks, and RQ4 compares this rate with model-dependent onset. Section 4.7 repeats the scoring without the opener check, which shows which conclusions depend on the opener. Format pass is not a general instruction-following score or a conversational usefulness rating. A reply may pass every hard check yet be factually inadequate, and a useful paraphrase may fail the opener.

Readability is reported as Flesch reading ease and Flesch–Kincaid grade [39,40]. Those formulae are validated on written text. They are not a listening-test substitute for synthesized speech, and no MOS or ITU-T P.85/P.808 listening scores were collected for this study [14,15].

3.10.2. Lexical Coverage, Echo, and Short-Span Presence

Lexical scoring reports IDF-weighted coverage of content words from the user utterance (surrogate; no reference required) and an echo ratio (verifiable). Coverage is computed twice when a scripted reference transcription is available: once against the recognized text (`stt_text`) and once against the intended question. Cov(ori) is lexical contact with intended wording; when recognition drops content words, that contact can fall for mechanical reasons as well as model behaviour. Format-compliant replies to a misheard utterance are therefore distinguished from contact with what the speaker asked (RQ5). Where the scenario spec supplies reference answers, token-level F1 [38] and ROUGE-1/ROUGE-L [41] are computed against the best-matching reference, and a presence indicator records whether the normalized gold span appears in the reply. Presence is reported on the 982 answerable items whose gold answer has at most eight words; longer gold spans are excluded from that denominator. Presence is an exact-span indicator, not a graded correctness score: a reply may contain the span yet be wrong elsewhere, and a correct paraphrase that omits the span is counted as a miss. Overlap metrics correlate weakly with human dialogue judgements [42] and are used as filters, not as quality scores.

3.10.3. Paired Configuration Contrasts

Paired comparison of two runs that answered the same inputs uses the user text as the match key, not list position. Binary metrics are tested with McNemar's test [43]; continuous metrics with the Wilcoxon signed-rank test [44]. Direction of improvement is declared per metric; length statistics remain descriptive. Holm adjustment [45] is applied to the family of scored metrics. Effect size is Cliff's δ [46]. Where a practical margin is predeclared, two one-sided tests may support equivalence rather than mere non-significance [47].

Run-level means use percentile bootstrap intervals [48]. Artefacts are `evaluation_report.txt`, `evaluation_items.csv`, and `evaluation_results.json`, written beside the pipeline run. Numerical prompt-adherence and coverage outcomes are reported in Section 4.

3.11. Experimental Protocol

A cell is a resolved model, decoding, recognizer-input, and synthesis configuration. A launch is a new execution of the assistant for that cell; a round contains one launch of each selected cell. Repeating recordings within a process and restarting the process are not equivalent forms of replication. The protocol uses rounds to measure variation across independently launched pipeline processes as well as variation across recordings.

The matrix runner accepts the semicolon-delimited configuration matrix through its `file-name` column, resolves the named YAML files, and launches them serially. Other columns describe the configurations but are not a second source of runtime overrides. For round r, the cell list is shuffled with a local random generator seeded by the matrix seed plus r. When more than one cell is present, the runner also prevents the same cell from appearing on both sides of a round boundary. This produces a reproducible interleaved cell order across rounds; the item order within a launch is not randomized, so every launch of a cell processes the items in the same order.

Each cell is started in a fresh assistant process. The Ollama server is not restarted for every cell, and cached model state can persist according to its retention policy. This protocol measures warm service behaviour under independent client launches, not fully cold independent servers. Fixed decoding settings and input order constrain the experiment while the shuffled cell order reduces systematic confounding with position in the matrix. Shared-host drift remains possible; the three-launch design tests whether it changed the scored replies under this warm serving stack (Section 4.10).

The model-size and decoding comparison uses 24 Vosk configurations and 12 Whisper configurations, each assigned three independent launches under Ollama 0.33.3. This gives 72 Vosk and 36 Whisper launches for that comparison set. The weight-precision comparison is defined by the nine model–precision combinations at temperature 0 in Section 3.4; these combinations overlap with the model inventory of the size/decoding comparison and are not added to it as nine independent repetitions. Configuration counts, launch counts, and matched-item counts are recorded separately for each analysis. The split-versus-full TTFA validation experiment has its own declared subset and repetition count. Before analysis, each launch is checked against the eligible filename list, audio–reference exclusions, transcript provenance, expected stages, and skip/failure markers; missingness is reported per arm and cell.

3.12. Statistical Analysis

The design separates two units of analysis. The item (recording, or its frozen transcript) is the unit for recognition fidelity, format pass, lexical scores, and item-paired configuration contrasts. The launch (independent client process for a cell) is the unit for host/serving replication and for launch-aware latency summaries. Recordings vary in speech duration, endpoint behaviour, recognized wording, and response content. Independent launches can vary with backend and host state even when the input and decoding settings are fixed. Treating all recording-by-launch rows as independent experiments would overstate replication. Likewise, a narrow interval over recordings in one launch cannot establish reproducibility across launches.

For latency, the summary report first computes the selected summary within each launch after first-item exclusion, then reports the median and range of those launch-

level summaries across rounds. For response metrics, it summarizes within-launch item means and then their across-launch mean and range. Raw item-level distributions remain available for the median and upper-tail views. Median-first reporting reflects the skew and consequential slow tail of interactive latency [49,50]. Quantiles are accompanied by their actual number of usable observations; extreme-tail precision is not inferred from a large nominal corpus when missingness or a small validation subset reduces the effective sample [51]. In the reported tables, factorial stage medians (TTFT, TTFC, first-chunk synthesis, generation, throughput) are taken from the per-launch latency aggregates written with each run; text-side quality rates and recognition-error strata are taken from the post hoc evaluation archive over the same transcripts. Device memory figures use the runner-reported LLM allocation in decimal gigabytes. Absolute speech-end–anchored TTFA is quoted only from the measured validation subset (Section 4.3).

Response-scoring comparisons in Section 3.10 use item-level analysis machinery: McNemar's test for binary contrasts [43], Wilcoxon signed-rank tests for continuous contrasts, 2000 bootstrap resamples for paired mean-difference intervals [48], and Holm adjustment within the stated metric families [45]. These item-based calculations characterize paired recordings and do not replace launch-level replication. The two analysis levels and their uncertainty estimates are kept distinct.

Primary contrasts and their metric families must be fixed before interpreting the evaluation; exploratory associations are labeled as such. Model size, quantization, recognition source, decoding temperature, and serving version are kept separate in interpreting differences. Failure to reject a difference is not evidence of equivalent response quality. Equivalence requires an explicit practically meaningful margin and an appropriate paired analysis [47]; neither a passing format check nor a small lexical-score difference establishes semantic interchangeability of quantized models.

For recognition robustness, intended-question and recognized-input coverage remain separate outcomes. Their association with WER within an arm is descriptive of naturally occurring strata, not a causal effect of an experimentally randomized error rate. Cross-arm Vosk–Whisper contrasts are compound regime contrasts (Section 3.3). No human preference, perceptual TTS quality, or application-level effectiveness is estimated from these automatic proxies.

3.13. Ethical, Privacy, and Safety Considerations

The study concerns technical performance and format pass on open-domain read-aloud questions. It contains no prospective user intervention, application-specific participant study, or assessment of user outcomes. Corpus selection and metadata preparation are engineering procedures; no expert suitability rating is implied.

Runtime recognition, generation, and synthesis take place on the evaluation host rather than through third-party speech or language APIs. This is a local-data-handling design choice, not a claim of certified privacy or regulatory compliance. Logs and retained waveforms can contain source questions, recognized text, and generated replies and require appropriate access control. The system prompt constrains format and encourages factual caution but is not a general safety policy. Domain-specific, adversarial, multi-turn, and human-assessed evaluation remain outside the present corpus and measurement protocol.

4. Results

This section reports the evaluation of Section 3.11 on the 999 surviving HeySQuAD recordings of Section 3.7 (two defective files excluded from the scored denominators; Table 4). Recognition and generation were split. A single Vosk pass and a single Whisper

pass produced frozen transcripts; the language-model factorial then answered those transcripts under a fixed Piper path [16]. The size and decoding comparison comprises 24 Vosk cells and 12 Whisper cells, each launched three times in interleaved round order, using the Llama 3.2/3.1 [52,53], Qwen2.5 [54], and Gemma 3 [55] tags of Table 3 served locally through Ollama over llama.cpp [18,19]. Weight precision was additionally varied at $T = 0$ on `llama3.1:8b`, `qwen2.5:7b`, and `gemma3:4b`. Direct speech-to-speech TTFA was measured on a randomly sampled 200-item subset (seed 42) for five cells, and reconstruction agreement was assessed against that measured arm (Section 3.9).

Every scored launch returned 999 non-empty replies with no skip or worker exception. Aggregate quality metrics (adherence, coverage, presence) are identical across the three launches of each cell (Section 4.10). Unless a launch range is stated, tables report launch r1 (quality metrics showed zero range across launches). Scoring used the verifiable-constraint and lexical instruments of Section 3.10. Format pass denotes a strict item-level hard-check conjunction [22,23], dominated by the five-word opener, rather than usefulness, factual correctness, or perceptual TTS quality [15,42]. Unless a caption states otherwise, adherence, coverage, and presence rates are item-level proportions on the stated denominator; stage times are launch-level medians (or medians of launch medians where noted); inferential contrasts that pair recordings are item-level. Item-paired contrasts use McNemar or Wilcoxon tests with Holm adjustment within each metric family and percentile-bootstrap 95% intervals [43–46,48]; equivalence is claimed only where two one-sided tests against a predeclared margin succeed [47].

4.1. Recognition Fidelity of the Shared Input Set

Because every downstream cell of a given arm answered the same recognizer output, recognition accuracy is a property of the input set and bounds what any model could have answered. Against the scripted reference, Vosk `small-en-us-0.15` [11,32] produced 2189 word errors over 10,926 reference words (corpus WER 0.200), composed of 1541 substitutions, 85 deletions, and 563 insertions: the dominant failure remains a misheard word rather than a lost one. The median item-level WER is 0.167, and 278 of 999 utterances (27.8%) were recognized exactly. Whisper `small` (CUDA, int8) [9] on the same recordings produced 523 errors over 10,772 reference words (corpus WER 0.049), with 693 exact transcripts (69.4%). The two denominators differ because each engine applies its own numeric and spoken-form normalizations (10,926 vs. 10,772 reference words). The fidelity contrast (0.200 vs. 0.049) stays large under either reading. Exact-transcript counts on the shared 999 recordings are 278 vs. 693, so the gap is not only a denominator effect. Table 6 defines the input-quality strata used below. The two arms are not pooled, and the cross-arm contrast remains compound (Section 3.3). WER alone does not determine understanding accuracy [24–26,37].

On the Vosk arm, 721 of 999 items present the model with a question that differs from the one spoken. The Whisper arm reduces that mismatch to 306 items and almost eliminates the severe stratum (22 vs. 213). A low coverage or presence figure on Vosk therefore cannot be attributed to the model without reading the recognized text alongside it.

The Vosk speech-end anchor is usable on every recording: endpoint delay is positive throughout, with median 975 ms (range 517–1709 ms; 95th percentile 1109 ms). Four of the 999 recordings produced two finalized Vosk results (an endpoint inside the utterance); their transcripts join both parts.

Table 6. Recognition fidelity of the shared input set (999 recordings; item-level strata). Word and character error rates are computed against the scripted reference, taking the best of several plausible spoken readings for numeric tokens. Content recall is the share of the reference's content words that survive recognition. Vosk is the streaming higher-error arm; Whisper is the buffered low-error arm.

Recognizer	Stratum	Items	WER, p50	CER, p50	Recall, p50
Vosk CPU	Clean (WER = 0)	278	0.000	0.000	1.00
	Mild ($0 < \text{WER} \leq 1/3$)	508	0.176	0.071	0.83
	Severe (WER $> 1/3$)	213	0.500	0.206	0.62
	All recordings	999	0.167	0.065	0.86
Whisper GPU	Clean (WER = 0)	693	0.000	0.000	1.00
	Mild ($0 < \text{WER} \leq 1/3$)	284	0.125	0.038	0.83
	Severe (WER $> 1/3$)	22	0.449	0.182	0.67
	All recordings	999	0.000	0.000	1.00

4.2. Component-Level Runtime

Table 7 reports the LLM and first-chunk synthesis stages of Table 5 on frozen Vosk transcripts at $T = 0$. These stages are the model-dependent part of onset. They are not TTFA: the shared Vosk endpoint (median 975 ms) and handover sit upstream, and the split design does not record a speech-end–anchored first-audio event on the factorial (Section 4.3).

Table 7. Per-stage runtime on frozen Vosk transcripts at $T = 0$, launch r1 (unit of analysis: launch-level stage medians over items), medians in milliseconds except throughput and device memory. TTFT = request to first token; TTFC = request to first speakable chunk; TTS_1 = synthesis of that chunk; Gen = server-side generation of all tokens; E2E = LLM-side response-ready span (no speech-end anchor). VRAM is the runner-reported LLM allocation in gigabytes.

Configuration	TTFT	TTFC	TTS_1	Gen	Tok/s	VRAM (GB)
`llama3.2:1b` Q4	29	88	180	60	198.5	1.27
`llama3.2:3b` Q4	39	142	219	120	131.5	2.54
`llama3.1:8b` Q4	48	168	176	392	89.9	5.17
`llama3.1:8b` Q8	52	220	196	534	68.0	8.37
`llama3.1:8b` F16	54	295	197	775	46.4	14.93
`qwen2.5:1.5b` Q4	40	127	209	94	157.8	1.44
`qwen2.5:3b` Q4	47	164	227	153	123.8	2.38
`qwen2.5:7b` Q4	52	170	180	179	93.1	4.90
`qwen2.5:32b` Q4	119	344	146	769	29.6	19.68
`gemma3:1b` Q4	80	186	236	187	128.1	1.23
`gemma3:4b` Q4	149	230	167	312	93.0	4.03
`gemma3:27b` Q4	657	857	141	1452	31.5	18.31

Throughput and memory order by parameter count and precision as expected. TTFC does not: nine of twelve cells reach a speakable chunk in 88–230 ms, `qwen2.5:32b` in 344 ms, and `gemma3:27b` in 857 ms. That last gap is almost entirely pre-token (657 ms TTFT). First-chunk synthesis is 141–236 ms and tracks opener length rather than size: the two largest models emit a five-word opener and are the fastest to synthesize it. Across the twelve $T = 0$ Vosk cells (launch r1), opener word count and first-chunk synthesis time correlate at Spearman $\rho = 0.58$–0.84 over items; replies whose opener has at most five words are synthesized in a median of 107–144 ms, against 181–245 ms for longer openers. Launch-to-launch medians for these stages move by at most a few milliseconds; item-level ICC$(2,1)$ for TTFT and throughput is poor [56,57], as expected for host jitter on short intervals, so configuration claims below use launch-level medians and item-paired tests, not item-level latency ICC.

4.3. End-to-End Speech-to-Speech Performance

Direct TTFA is available on the 200-item validation subset for five cells, each launched three times under real-time Vosk pacing. Table 8 reports launch r1 medians; r2 and r3 agree to within 12 ms on the TTFA median. Perceived wait is 1282–1539 ms for four cells and 2025 ms for `gemma3:27b`. The 95th percentile lies 271–470 ms above the median. Figure 2 localizes the wait: the shared endpoint delay (965–968 ms) accounts for 48% of TTFA on `gemma3:27b` and 63–75% on the other four cells. The model enters the wait through TTFC; first-chunk synthesis is secondary.

Table 8. Measured TTFA on the 200-item validation subset (Vosk, $T = 0$, launch r1; TTFA columns are item medians within that launch). Bland–Altman agreement uses item-level medians across launches within each arm before pairing ($n = 199$ paired items per cell after first-item exclusion; 995 pairs overall) [58]. Bias is mean (measured − reconstructed); positive bias means the reconstruction underestimates wait. LoA = 95% limits of agreement. Endpoint-dominance claims are for the calibrated 600 ms Vosk silence protocol.

Configuration	TTFA	p95	Bias	LoA	≤100 ms
`llama3.2:1b` Q4	1282	1553	+58	−40, +157	87.4%
`gemma3:1b` Q4	1434	1675	+63	−43, +169	83.4%
`qwen2.5:32b` Q4	1508	1842	+53	−45, +152	91.0%
`llama3.1:8b` F16	1539	2009	+41	−60, +143	94.0%
`gemma3:27b` Q4	2025	2245	+37	−48, +122	96.5%
Overall (995 pairs)	—	—	+50.6	−49, +150	90.5%

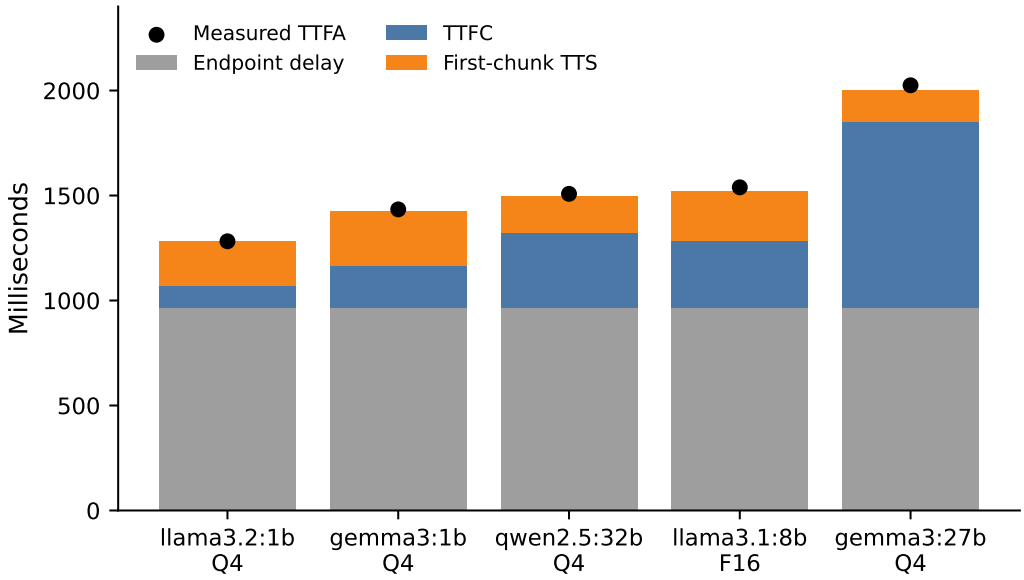


Figure 2. Measured TTFA (markers) and stage stack on the 200-item validation subset, launch r1. Endpoint delay is shared to within 3 ms. The `gemma3:27b` outlier is a pre-token interval, not a slower endpointer or a longer first speech chunk.

Item-matched reconstruction, using the ASR endpoint of the split pass plus downstream TTFC and first-chunk synthesis plus a 5 ms handover, correlates with measured TTFA at Spearman $\rho = 0.977$ [59]. The signed bias is +50.6 ms (reconstruction underestimates), with MAE 56.6 ms. The 95% limits of agreement (−49 to +150 ms) cross the predeclared 100 ms absolute-wait margin that had been set for a planned TOST equivalence test. Bland–Altman agreement and average equivalence are distinct claims [47,58]; crossing the LoA bound alone is not a TOST result. Reconstructed TTFA therefore supports stage

decomposition and configuration ranking on this host, not substitution for absolute wait. Absolute TTFA levels in this subsection are therefore the measured validation figures; the factorial contributes the stage breakdown of Table 7.

How strongly the endpoint dominates depends on the endpointer setting (Appendix B.2). In an offline sweep on the evaluation recordings the median endpoint delay follows the flat silence threshold almost one-to-one, from 0.66 s at 300 ms to 1.14 s at 800 ms, while premature mid-sentence splits rise from 0.34% at 600 ms to 1.70% at 400 ms and 4.86% at 300 ms. The rest of the wait (LLM time to first chunk, first-chunk synthesis, and handover) begins only after the endpoint (Equation (2)), so on recordings that are not split it does not depend on this setting. On the validation subset its median is 0.33–0.54 s for the four cells other than `gemma3:27b` and 1.04 s for `gemma3:27b`. For those four cells the endpoint therefore remains the larger part of the wait at every tested threshold, whereas for `gemma3:27b` it does so only above about 700 ms. The other factors have smaller effects: chunk length moves the endpoint delay by about 90 ms between 50 and 250 ms, the appended trailing silence did not shorten the measured endpoint, and a ± 100 ms shift of the speech-end anchor changes the endpoint share by at most three percentage points.

The wall-clock item span on the validation subset remains dominated by real-time delivery of the recording (median E2E 7.7–8.0 s) and is not a conversational latency [49]. Whisper's median processing interval after stream completion is 256 ms.

4.4. Effect of Weight Precision

Weight precision was varied within three architectures at $T = 0$, with F16 as the within-family baseline (Table 9). Memory and throughput move strongly and in the expected direction. Response effects do not share a single sign, so no universal "quantization does not hurt quality" claim is licensed [60].

Table 9. Weight-precision contrasts at $T = 0$ on frozen Vosk transcripts, launch r1. Adher. and Pres. are item-level rates; throughput and VRAM are launch-level medians. Adher. = strict item-level hard-check conjunction (opener-dominated); Pres. = exact-span presence on 982 answerable items with gold answers of at most eight words (question-only prompt; no passage). Q4/Q8 rows are compared with the F16 cell of the same family, not across families.

Family	Prec.	Adher.	Pres.	Tok/s	VRAM (GB)
Llama 8B	F16	17.8%	12.9%	46.4	14.93
	Q8_0	18.3%	12.9%	68.0	8.37
	Q4_K_M	25.9%	12.2%	89.9	5.17
Qwen 7B	F16	14.5%	8.1%	48.2	14.06
	Q8_0	15.1%	8.1%	73.3	7.89
	Q4_K_M	14.7%	7.4%	93.1	4.90
Gemma 4B	F16	45.5%	8.7%	65.7	9.02
	Q8_0	45.6%	8.6%	83.8	5.59
	Q4_K_M	39.3%	9.7%	93.0	4.03

On `llama3.1:8b`, Q4 versus F16 raises item adherence by 8.1 pp (95% interval 5.7–10.6; McNemar $p_{\text{Holm}} < 10^{-8}$) while lowering recognized-question coverage (mean shift -0.036) and leaving answer presence unchanged. The adherence gain coincides with a shorter opener (median 7 vs. 9 words), which the five-word rule rewards. Q8 is equivalent to F16 on check-level rate, coverage, echo, and presence within the 0.02 TOST margin. Device allocation falls from 14.93 to 5.17 GB (factor 2.9); throughput rises from 46.4 to 89.9 tokens/s; TTFC falls by 127 ms at the median.

On `qwen2.5:7b`, item adherence does not change detectably after Holm adjustment for Q8 versus F16 ($+0.6$ pp; interval -0.5 to $+1.6$) or Q4 versus F16 ($+0.2$ pp; interval

−2.1 to +2.5). Check-level pass rate, coverage, echo, and presence stay within the 0.02 TOST margin for those contrasts. For this tag the precision ladder mainly changes memory and throughput on the scored instruments. On `gemma3:4b`, Q4 lowers item adherence relative to F16 by 6.2 pp (interval 3.1–9.2; Holm-adjusted McNemar $p = 0.002$). Q8 shows no detected item-adherence change versus F16 (+0.1 pp; interval −1.1 to +1.2). Four-bit compression is not format-neutral on this family and this specification. All three family statements are limited to these tags, this host, and the instruments actually scored.

4.5. Configuration Size, Decoding Temperature, and Associated Outcomes

Three Q4 ladders are measurable among the tested tags: Llama at 1, 3, and 8 B; Qwen2.5 at 1.5, 3, 7, and 32 B; Gemma 3 at 1, 4, and 27 B (Table 10). These are model-configuration comparisons, not isolated estimates of parameter count (Section 3.4). Among the tested tags, larger Qwen and Gemma cells associate with higher format pass and lower throughput; the Llama ladder, which mixes Llama 3.2 and Llama 3.1 tags and tops out at 8 B, does not order format pass by size alone.

Table 10. Configuration ladders at Q4_K_M and $T = 0$ on frozen Vosk transcripts, launch r1. Adher. = item-level strict hard-check conjunction (opener-dominated). TTFC is the launch-level median request-to-first-chunk time.

Family	Ends	Adher. (Small)	Adher. (Large)	TTFC (Small)	TTFC (Large)
Llama	1 B → 8 B	16.7%	25.9%	88	168
Qwen2.5	1.5 B → 32 B	6.0%	66.8%	127	344
Gemma 3	1 B → 27 B	8.3%	70.8%	186	857

In the Qwen ladder at $T = 0$, stepping from 32 B to 7 B costs 52.1 pp of item adherence (95% interval 48.6–55.4; Holm-adjusted McNemar $p < 10^{-100}$); the 1.5 B and 3 B cells remain at 6.0% and 1.9%. Coverage of the recognized question does not follow that collapse, so the small Qwen cells fail the opener constraint while still echoing the question. In the Gemma ladder, adherence rises from 8.3% (1 B) through 39.3% (4 B) to 70.8% (27 B). In the Llama ladder, which tops out at 8 B, size does not order format pass: `llama3.2:3b` (10.7%) underperforms both the 1 B (16.7%) and 8 B (25.9%) cells. On the 1 B cell the median opener over all items is nine words (Table 11), so most replies fail the five-word hard check; the 16.7% that pass still show the lowest intended-question coverage among the Llama cells (0.420). Below approximately 8 B, adherence and content contact are therefore not two readings of one latent scale.

TTFC barely moves until the two largest models: the step from Llama 8 B Q4 to 1 B saves 80 ms of TTFC, and from Qwen 32 B to 1.5 B saves 217 ms, against several hundred milliseconds of generation time. First-chunk synthesis offsets part of that gain, because smaller models emit longer openers. The conversational onset is dominated by the shared endpoint plus, for `gemma3:27b` only, a long pre-token interval (Section 4.3).

Decoding temperature is a weak factor on the applied instruments under the fixed seed used here. For `gemma3:27b`, $T = 0.7$ (seed 42) versus $T = 0$ changes item adherence by +1.2 pp (interval −1.3 to +3.7; not significant) and is equivalent on check-level rate, coverage, and echo within the 0.02 margin. Across the twelve decoding pairs ($T = 0$ versus $T = 0.7$ at seed 42), the largest temperature shift is +5.6 pp on `llama3.2:1b` (16.7% to 22.3%), still far smaller than the 27–32 B separation. These contrasts compare greedy decoding to one sampled realization per item at seed 42; they do not estimate seed-level sampling variance.

Table 11. Spoken-format and lexical measures on frozen Vosk transcripts at $T = 0$, launch r1 (item-level rates; $n = 999$ items; presence on 982 short-span items, gold answer ≤ 8 words; question-only prompt). Adher. = strict item-level hard-check conjunction (opener-dominated; not conversational usefulness); Chk. = mean share of hard checks passed; Cov(stt)/Cov(ori) = median coverage of the recognized and intended questions; Pres. = exact-span presence indicator; Open = median opening-sentence word count.

Configuration	Adher.	Chk.	Cov(stt)	Cov(ori)	Pres.	Open
`llama3.2:1b` Q4	16.7%	90.1%	0.465	0.420	3.7%	9
`llama3.2:3b` Q4	10.7%	90.5%	0.624	0.526	6.8%	10
`llama3.1:8b` Q4	25.9%	92.5%	0.568	0.542	12.2%	7
`qwen2.5:1.5b` Q4	6.0%	89.3%	0.726	0.640	6.7%	10
`qwen2.5:3b` Q4	1.9%	89.9%	0.704	0.653	8.7%	11
`qwen2.5:7b` Q4	14.7%	90.3%	0.514	0.448	7.4%	9
`qwen2.5:32b` Q4	66.8%	96.5%	0.515	0.470	12.9%	5
`gemma3:1b` Q4	8.3%	90.5%	0.656	0.587	5.8%	11
`gemma3:4b` Q4	39.3%	93.9%	0.601	0.547	9.7%	6
`gemma3:27b` Q4	70.8%	96.9%	0.599	0.544	13.1%	5

4.6. Verifiable Spoken-Format Constraints and Lexical Surrogates

Table 11 reports the spoken-format and lexical measures. Format pass is the strict conjunction of the hard checks of the short-opener specification (Section 3.6), dominated by the five-word opening sentence. The check-level rate remains high (89–97%) even where item pass is rare, because most checks (English, non-empty, no markup, no simulated dialogue) almost always hold [22,23]. Presence is the share of 982 answerable items whose gold answer has at most eight words and whose reply contains that gold span under the scorer's SQuAD-style normalisation (lowercase, punctuation stripping, and article removal) [38]. It is an exact-span indicator, not a graded quality score. Exact match is near zero by construction, because replies are sentences rather than spans.

Only the two largest models meet the five-word opener at the median and pass the conjunction on a majority of items. Their presence rates remain 13.1% and 12.9%: format compliance does not imply exact-span presence, and a correct paraphrase that omits the span is counted as a miss. Because generation is question-only (Table 4), these rates also reflect items whose gold spans are recoverable mainly from the omitted SQuAD passage; they are not passage-grounded reading-comprehension scores. Section 4.7 stratifies omitted-context dependence and rescores the hard-check conjunction without the opener.

Small models often cover more of the recognized question by echoing it (echo medians 0.36–0.60 versus 0.18 for `gemma3:27b`). Token-level F1 against the gold span is 0.045–0.074 and is depressed by every non-span word of the spoken sentence; it is reported as a filter, not as a dialogue-quality estimate [42]. Readability formulae [39] place the replies at grade 7–9; they are not a substitute for a listening test.

Paired comparison of the two largest cells at $T = 0$ does not establish an adherence ranking after Holm adjustment (+4.0 pp for `gemma3:27b` versus `qwen2.5:32b`; interval 0.2–7.6; $p_{\text{Holm}} = 0.55$). Gemma covers more of the recognized and intended questions and echoes less; Qwen is 498 ms faster to the first speakable chunk (interval 490–505) and 1.27 s faster on the LLM-side response-ready span. The choice between them is a wait-versus-verbosity trade, not a validated quality ranking.

4.7. Sensitivity to Omitted Context and Opener Ablation

The primary tables above use the full hard-check conjunction and the full short-span presence denominator ($n = 982$). Two post hoc rescores of the same archived replies bound

how those tables may be read. Neither regenerates model outputs with SQuAD context in the prompt, and neither replaces the opener with a semantic task-achievement criterion.

4.7.1. Context Strata

Each of the 999 items was labelled from the question string and gold answers after the same SQuAD-style normalisation used for presence. These labels are lexical proxies, not validated measures of question-only answerability: *zero gold–question overlap* if every gold answer has zero content-token overlap with the question ($n = 758$); *positive gold–question overlap* otherwise ($n = 241$); *referential* if the question contains overt discourse deixis ($n = 78$); and *overlap without overt deixis* if positive overlap and not referential ($n = 228$). Short-span presence on that restricted set uses $n = 220$ items (gold answer ≤ 8 words). Table 12 reports the two largest Vosk cells at $T = 0$ and their Whisper counterparts.

Table 12. Post-hoc sensitivity on archived replies (item-level rates; no new generation). Full format pass is the primary hard-check conjunction; opener-ablated format pass drops only `opener_at_most_five_words` from the failure set. Short-span presence uses gold answers of at most eight words. Pres. (alone) uses the lexical proxy *overlap without overt deixis* (positive gold–question overlap and not referential); it is not a validated exclusion of context-dependent questions. Rescored from the archived evaluation tables.

Configuration	Format	Opener-Abl.	Pres. (Full)	Pres. (Alone)
`gemma3:27b` Vosk	70.8%	98.3%	13.1% (982)	18.2% (220)
`qwen2.5:32b` Vosk	66.8%	97.1%	12.9% (982)	15.5% (220)
`gemma3:27b` Whisper	68.4%	99.7%	17.1% (982)	19.5% (220)
`qwen2.5:32b` Whisper	72.5%	98.2%	17.9% (982)	18.2% (220)

On Vosk, short-span presence rises from 13.1%/12.9% on the full short-span set to 18.2%/15.5% on the overlap-without-overt-deixis subset, and remains 11.7%/12.3% on the zero-overlap short-span subset ($n = 749$). The rate increases on this lexically restricted subset, but stays well below a majority. These proxies do not establish performance after excluding all context-dependent or underdetermined questions, and they do not turn exact-span presence into a strong correctness claim under the question-only prompt. No replies were regenerated with the SQuAD passage in the prompt.

4.7.2. Opener Ablation

Rescoring the hard-check conjunction after removing the five-word opener failure raises format pass to 98.3% and 97.1% on the two large Vosk cells (opener-only fails 27.5% and 30.3% of items). Across all twelve temperature-0 Vosk factorial cells, opener-ablated format pass is at least 50% on every cell (range 88.8–99.6%), whereas majority pass under the full conjunction remains confined to the 27–32 B cells. Mid-size cells show the opener share: `llama3.1:8b` Q4 rises from 25.9% to 99.2% (opener-only fails 73.3%), and `gemma3:4b` Q4 from 39.3% to 98.8% (opener-only fails 59.5%). The size-gated majority-pass pattern in the primary tables is opener-dominated. It does not hold for the remaining hard checks alone. Under the full conjunction (with opener), majority pass occurs only at 27–32 B, and opener failures account for most full-conjunction misses on mid-size cells. Two other format-pass claims also rest on the opener. Without that check, format pass is 99.6, 99.6, and 99.2% for Llama 8B at F16, Q8, and Q4, and 98.4, 98.6, and 98.8% for Gemma 4B; the Q4 gain on Llama and the Q4 loss on Gemma (Section 4.4) do not survive. Qwen 7B moves from 85.8% (F16) to 88.8% (Q4). On Whisper transcripts the five models reach 94.8–99.7% opener-ablated format pass, at or above their Vosk values, so the cross-arm adherence shifts are opener effects too. Across the 24 Vosk cells, opener-ablated format pass spans 73.5–99.6%, and the 27–32 B cells (96.7–98.5%) are no longer the highest: the front in Figure 3

is specific to the full conjunction. TTFA, endpoint share, memory, throughput, stage latency, WER, coverage, and presence do not use the format instrument, so the ablation does not change them.

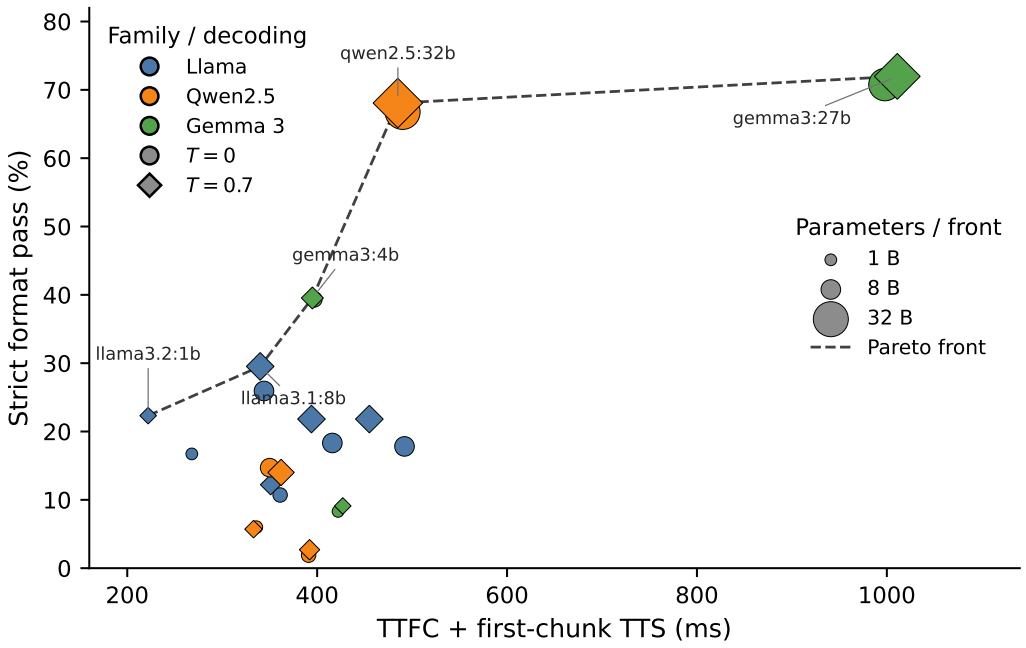


Figure 3. Model-dependent onset (TTFC + first-chunk TTS; launch-level medians) against item-level strict format pass for the 24 Vosk cells, launch r1. Marker area scales with parameter count; colour encodes family; circles $T = 0$, diamonds $T = 0.7$. The dashed line is the non-dominated front under a 50 ms wait time: `llama3.2:1b`, `llama3.1:8b` Q4, `gemma3:4b`, `qwen2.5:32b`, and `gemma3:27b`, all at $T = 0.7$. Front membership is descriptive under opener-dominated format pass only. It is not a temperature recommendation or a ranking of conversational usefulness. Along that front, format pass rises from 22.3% to 68.1% for 263 ms of additional onset, then only to 72.0% for a further 526 ms.

4.8. Compound Recognizer Regimes and Within-Arm WER Strata

RQ5 uses two compound recognizer regimes on the same recordings rather than injected token noise under a fixed recognizer. Within the Vosk arm, Table 13 stratifies four cells by naturally occurring WER. Presence and intended-question coverage fall from clean to severe for every model. Item adherence does not: `gemma3:27b` is stable across strata (70.5–71.4%), `qwen2.5:32b` drops from 72.7% to 62.0%, and `gemma3:4b` rises on severe items, consistent with shorter, more damaged inputs making the opener rule easier to satisfy. Under these instruments and WER strata, lexical contact with the intended question (Cov(ori) and short-span presence) falls more than format pass. Much of that split is built into the scores: format pass is dominated by the five-word opener, while Cov(ori) falls when recognition drops content words that an echoing reply cannot restore. Stratum membership also co-varies with item properties [24,37]. Spearman correlations of item WER with intended-question coverage are negative for all four cells ($\rho = -0.22$ to -0.37); the correlation with adherence is near zero for `gemma3:27b` ($\rho = -0.001$) and only weakly negative for the others. These within-arm associations are descriptive, not a randomized error-rate experiment.

On the same recordings Whisper lowers corpus WER from 0.200 to 0.049, and the downstream factorial was re-run on its transcripts (12 cells, three launches each). Repeating the Q4 factorial on Whisper transcripts raises gold-span presence for every listed model (Table 14). Adherence moves in both directions: `llama3.1:8b` gains 13.2 pp and `qwen2.5:32b` gains 5.7 pp, while `gemma3:27b` loses 2.4 pp and `gemma3:4b` loses 7.3 pp.

The Whisper arm is the low-error regime (corpus WER: 0.049), but the Vosk–Whisper contrast remains compound: architecture, decoding behaviour, transcript timing, and error type change together with WER (Section 3.3). A lower-WER transcript is not a pure error-propagation treatment and does not uniformly raise format pass. The Whisper severe stratum is too small ($n = 22$) for stratum-wise inference; the arm-level contrast is the intended robustness result.

Table 13. Response measures by Vosk input-quality stratum at $T = 0$, launch r1 (item-level rates within strata). Strata as in Table 6. Presence is the exact-span indicator within the stratum (short-span gold answers; question-only prompt).

Configuration	Stratum	Adher. (%)	Cov(ori)	Pres. (%)
`llama3.1:8b` Q4	clean	30.6	0.68	15.5
	mild	26.2	0.55	12.8
	severe	19.3	0.34	5.6
`gemma3:4b` Q4	clean	34.9	0.68	12.9
	mild	40.0	0.54	10.0
	severe	43.7	0.39	3.8
`gemma3:27b` Q4	clean	70.5	0.63	13.3
	mild	70.7	0.54	14.6
	severe	71.4	0.45	8.5
`qwen2.5:32b` Q4	clean	72.7	0.58	16.2
	mild	65.6	0.50	13.6
	severe	62.0	0.31	6.1

Table 14. Same Q4 models at $T = 0$ on frozen Vosk versus Whisper transcripts, launch r1 (item-level rates). Presence on 982 short-span items under the question-only prompt. Cross-arm differences are compound regime contrasts, not pure WER isolations.

Configuration	Adher. (Vosk)	Adher. (Whisper)	Pres. (Vosk)	Pres. (Whisper)
`llama3.1:8b` Q4	25.9%	39.1%	12.2%	18.1%
`qwen2.5:7b` Q4	14.7%	18.3%	7.4%	11.6%
`gemma3:4b` Q4	39.3%	32.0%	9.7%	12.9%
`qwen2.5:32b` Q4	66.8%	72.5%	12.9%	17.9%
`gemma3:27b` Q4	70.8%	68.4%	13.1%	17.1%

4.9. Latency–Format Trade-Off

Figure 3 places every Vosk cell in the plane of model-dependent onset (TTFC plus first-chunk TTS) and opener-dominated format pass. The shared endpoint is omitted so that the abscissa isolates the quantity the LLM actually controls. Marker area is proportional to parameter count (legend: 1, 8, and 32 B); colour encodes family; circles are $T = 0$ and diamonds $T = 0.7$. The dashed line is the Pareto front: no other cell is better on both axes, treating differences below 50 ms on the abscissa as ties.

Five cells are non-dominated. From `llama3.2:1b` to `qwen2.5:32b` the front is steep: 263 ms of extra model-dependent onset (222 to 485 ms) corresponds to 45.8 pp of format pass. The last segment is flat: `gemma3:27b` improves adherence by only 3.9 pp over `qwen2.5:32b` at a 526 ms TTFC premium, matching the ≈0.5 s measured-TTFA gap on the validation subset (Table 8). Intermediate front members (`llama3.1:8b` Q4 at 29.5%, `gemma3:4b` at 39.5%) improve on the 1 B cell without approaching the 32 B format-pass rate. Off-front cells, including the greedy partners of the five front members, are dominated under the 50 ms rule. Absolute speech-end–anchored waits for these cells are the measured validation figures in Table 8. Item-matched reconstruction on that subset has bias +50.6 ms and 95% limits of agreement −49 to +150 ms (Section 4.3). Reconstructed

totals are not reported as absolute onsets. All five front members are $T = 0.7$ cells from one seed-42 realization. The front is a descriptive single-seed sensitivity under opener-dominated format pass, not a seed-stable ranking. Exploratory Spearman correlations across the 24 launch r1 cells give $\rho = +0.68$ between adherence and parameter count, $\rho = +0.73$ with VRAM, $\rho = -0.68$ with throughput, and $\rho = +0.64$ with TTFC; these rest on one summary per cell, were not predeclared, and are unadjusted.

The front members are descriptive under opener-dominated format pass. They order cells only on model-dependent onset versus that hard-check rate. They are not a ranking of conversational usefulness, factual adequacy, or listening quality.

- Lowest onset on the front. `llama3.2:1b` at Q4 and $T = 0.7$ (222 ms model-dependent onset; 22.3% adherence). Its greedy partner is dominated under the 50 ms tie.
- End of the steep segment. `qwen2.5:32b` at Q4 and $T = 0.7$ (485 ms; 68.1% adherence). Its $T = 0$ measured TTFA on the validation subset is 1.51 s.
- Flat-front tail. `gemma3:27b` at Q4 and $T = 0.7$ (1011 ms; 72.0% adherence). Its $T = 0$ measured TTFA is 2.03 s. After Holm adjustment it has no adherence advantage over Qwen 32B on the hard-check conjunction (Section 4.6).
- Intermediate front member. `gemma3:4b` at Q4 and $T = 0.7$ (395 ms; 39.5% adherence), well below the 27 B pre-token cost, with presence still under 10% on Vosk.

4.10. Validity of the Reported Figures

Six properties bound what the preceding figures support, independently of the broader limitations in Section 5, following current reporting guidance for language-model evaluations [60].

Each configuration cell was launched three times in interleaved round order. Aggregate quality metrics (adherence, coverage, presence) are identical across r1–r3 for all 24 Vosk cells and all 12 Whisper cells, including the sampled ($T = 0.7$, seed 42) cells; ICC$(2,1)$ on item-level adherence is 1.0 [56,57]. Under this stack the inputs were frozen transcripts, decoding used temperature 0 or temperature 0.7 with seed 42, and the Ollama server remained warm across client relaunches. Identical aggregates are expected when the backend returns the same tokens for the same request; the raw replies (`llm_text`) are byte-identical item by item in every cell, including the five measured-TTFA cells with live Vosk recognition, and the controls do not guarantee identical outputs if host load, cache state, or serving software drifts (Section 3.11). The three launches are a host and serving determinism check, not an estimate of seed-level sampling variance for $T = 0.7$. Response-side inference remains item-level, conditional on one generation per item. Runtime medians are stable to a few milliseconds across launches; item-level latency ICC remains poor, so small TTFT differences are not interpreted.

Absolute wait is reported from the directly measured validation arm; the factorial supplies stage timings (Section 4.3).

Levels and shifts use different estimators. Tables report medians; paired contrasts estimate mean shifts with bootstrap intervals. Claims about differences rest on the contrasts; claims about levels rest on the descriptive tables.

The response measures are verifiable checks and lexical surrogates. Nothing above shows that a reply is helpful, factually correct, or acceptable as speech. Answer presence is an exact-span indicator on short gold answers. The open-domain judge protocol was prepared on a stratified 100-item sample but was not scored, because no local judge model was available at evaluation time; no rubric mean is reported. No MOS or ITU-T P.808 listening score was collected [15].

Quantization results are within-family. Q4 raised format pass on Llama 8B, left Qwen 7B item adherence with no detected change (check-level rate equivalent within the 0.02 margin), and lowered format pass on Gemma 4B (Section 4.4).

Finally, the input set comprises 999 English read-aloud factual questions without dialogue history. It exercises streaming endpoint behaviour, split-design stage timing, and format pass under two recognizer error regimes.

5. Discussion

5.1. Principal Findings

Every scored launch returned 999 non-empty replies (Section 4). On RQ1, TTFA on the measured validation subset is 1282–1539 ms for four cells and 2025 ms for `gemma3:27b` (Table 8). Shared Vosk endpoint delay (965–975 ms) accounts for most of that wait except on `gemma3:27b`, where a long pre-token interval adds about half a second (Figure 2). For the other four cells the endpoint remains the larger part of the wait at every flat silence threshold from 300 to 800 ms (Section 4.3).

On RQ2, among the tested tags, weight precision moves device memory and throughput strongly (Table 9). Under the full hard-check conjunction, Llama 8 B gains format pass under Q4, Qwen 7 B shows no detected item-adherence change across precisions with check-level equivalence within the 0.02 margin, and Gemma 4 B loses format pass under Q4 relative to F16. Opener ablation removes those family differences (Section 4.7). Larger Qwen and Gemma configurations are associated with higher format pass under the full conjunction (Table 10). Conversational onset barely moves until `gemma3:27b`. Its median TTFT is about 657 ms and its logged prompt-evaluation median about 640 ms, against about 101 ms prompt-eval and 119 ms TTFT for `qwen2.5:32b` at similar VRAM under this Ollama path. Decoding temperature at seed 42 remains the weakest factor. CPU/RSS/GPU-utilization samples were logged but not analysed as cell-level outcomes (Section 3.8).

On RQ3–RQ4, opener-dominated format-pass rates at $T = 0$ span 1.9–70.8% and rise mainly at 27–32 B under the full conjunction (Table 11). We keep that full conjunction, including the opener, as the primary format outcome. Post hoc opener ablation lifts every temperature-0 Vosk cell to 88.8–99.6% format pass, so the size gate is opener-specific and does not describe the remaining hard checks (Section 4.7). Gold-span presence stays low even on the overlap-without-overt-deixis lexical subset (15.5–18.2% for the largest Vosk cells). The wait–format plane (Figure 3) orders cells on model-dependent onset versus that hard-check rate only. Under that conjunction, `qwen2.5:32b` at Q4 sits at the end of the steep segment of the front, not as a usefulness ranking.

On RQ5, within the observed higher-error streaming Vosk condition (WER 0.200), lexical contact with the intended question falls from clean to severe more than format pass does (Section 4.8). Format pass is opener-dominated; Cov(ori) tracks content words that recognition may drop. The Whisper low-error regime (WER 0.049) raises presence but moves format pass in both directions. The cross-arm contrast is compound, not a pure error-propagation treatment, and no absolute TTFA is reported on Whisper. Format pass alone cannot detect that the exchange concerns the wrong subject [24,37].

5.2. Limitations and Scope

The findings are warm-protocol completion, stage-resolved latency, and text-side hard-check and lexical scores on English read-aloud, question-only items. They do not estimate application-level effectiveness, safety, helpfulness, multi-turn consistency, or perceptual TTS quality. No therapists, clinicians, or patients took part, and no therapeutic or clinical use is evaluated. No MOS, preference, or ITU-T listening test was collected; synthesis

timing is not a substitute for perceived speech quality. Absolute TTFA statements are restricted to the calibrated flat 600 ms Vosk silence protocol and the measured validation subset (Section 4.3). At other thresholds and chunk lengths only the endpoint delay was measured, in offline sweeps (Appendix B.2). Quantization findings are within-family. Three interleaved launches show byte-identical replies under the serving conditions of Section 3.4; they do not estimate seed-level sampling variance and do not guarantee determinism on other hosts. Item-level latency ICC remains poor. Exact-span presence is interpreted under omitted SQuAD context (Table 4); Section 4.7 applies lexical-overlap proxies and rescores opener ablation on archived replies only, without validated question-only answerability labels, context-in-prompt regeneration, or a semantic task-achievement criterion. A higher-accuracy streaming recognizer (including a Vosk upgrade) with a full factorial re-run, barge-in under spontaneous speech, and controlled domain evaluation remain future work.

5.3. Practical Deployment Considerations

On this evaluation host the largest four-bit cells allocate about 56–60% of the 32 GB accelerator, whereas `llama3.1:8b` at Q4 occupies 5.17 GB (Tables 7 and 9). No cell failed to fit. Device memory is reported as allocation, not as a constraint that ruled out configurations. Local serving keeps audio and weights on-host without claiming a regulatory mandate for on-premises inference [2]. Under the flat 600 ms Vosk silence protocol the endpoint is large, so silence handling should be calibrated for the intended speech setting; read-aloud timing does not transfer directly to spontaneous interaction. These results do not rank cells for deployment. Format pass scores surface rules of one prompt. Application choice would need domain answer-quality measures that this study does not provide.

6. Conclusions

This paper evaluated a locally served ASR–LLM–TTS cascade on 999 English read-aloud HeySQuAD questions, with split recognition and generation, two recognizer arms, three interleaved launches per cell, and a measured-TTFA validation.

Measured TTFA medians were 1.28–1.54 s for four validation cells and 2.03 s for `gemma3:27b` under the calibrated streaming Vosk path. Absolute TTFA is not reported for the buffered Whisper arm.

Four-bit compression reduced device memory by about three on Llama 8B and Qwen 7B and by about 2.2 on Gemma 4B. Format-pass effects differed by family: Llama 8B gained under Q4, Qwen 7B item adherence showed no detected change with check-level equivalence within the 0.02 margin, and Gemma 4B lost under Q4. Opener ablation removes both precision effects (Section 4.7). Among the tested tags, larger Qwen and Gemma configurations are associated with higher format pass and lower throughput under the full conjunction, but only `gemma3:27b` moved onset by about half a second. Under the full hard-check conjunction, format pass rose to a majority only at 27–32 B. That size gate is opener-specific: after removing the five-word opener check, all twelve temperature-zero Vosk configurations passed on 88.8–99.6% of items. Smaller models could pass the full conjunction while showing low intended-question coverage across all items. Gold-span presence on short-answer items remained low (13–18% even for the largest models). Absolute TTFA and endpoint-share statements refer to the calibrated flat 600 ms Vosk silence protocol on the measured validation cells (Section 4.3).

Under the observed higher-error streaming Vosk condition (WER 0.200), Gemma 27B format pass stayed intact while intended-question coverage and presence fell with WER strata. Format pass is opener-dominated; Cov(ori) tracks content-word contact. The Whisper low-error regime (WER 0.049) raised presence but did not uniformly raise format pass.

The cross-arm contrast remains compound. Aggregate quality metrics were identical across the three launches of each cell under fixed decoding settings, frozen inputs, and warm serving. Every warm turn of the protocol completed with a non-empty reply on this question-only corpus and these recognizer regimes.

Author Contributions: Conceptualization, B.G., K.E., Á.P., Z.B. and G.K.; methodology, B.G., K.E., Á.P. and G.K.; software, B.G., K.E. and Á.P.; validation, B.G., K.E., Á.P. and G.K.; formal analysis, B.G., K.E., Á.P., Z.B. and G.K.; investigation, B.G., K.E., Á.P., Z.B. and G.K.; resources, B.G. and G.K.; data curation, B.G. and K.E.; writing—original draft preparation, B.G., K.E., Á.P. and G.K.; writing—review and editing, B.G., K.E., Á.P., Z.B. and G.K.; visualization, B.G. and Á.P.; supervision, G.K.; project administration, G.K.; funding acquisition, G.K. All authors have read and agreed to the published version of the manuscript.

Funding: Project no. 2023-1.2.1-ERA_NET-2023-00017 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2023-1.2.1-ERA_NET funding scheme.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Per-cell run artefacts (`config_used.yaml` with resolved tags and paths, system-prompt copies, latency CSVs, transcripts) and post hoc evaluation aggregates (leaderboards, launch-replicate summaries, paired contrasts, TTFA reconstruction agreement, summary reports) are archived with the accompanying materials at https://github.com/oe-etap/etap-speech-assistant (accessed on 10 August 2026). Each analysed cell is identified by its archived `cell_id` and launch tree. Mutable Ollama tags alone are not the experiment record. The scored aggregates cover the $n = 999$ validated recordings; two defective files are excluded as documented in Table 4. Source audio waveforms are obtained from the HeySQuAD spoken question-answering release [37] and are not redistributed here; item identifiers accompany the evaluation archive. Constraint specifications and analysis scripts for the verifiable-constraint and lexical instruments accompany the same archive.

Acknowledgments: On behalf of the "Distributed deep learning in parallel and distributed environment" project, we are grateful for the possibility of using HUN-REN Cloud (see [31]; https://science-cloud.hu/ (accessed on 10 August 2026)), which helped us achieve the results published in this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
E2E	End-to-End
GGUF	GPT-Generated Unified Format
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HTTP	Hypertext Transfer Protocol
IDF	Inverse Document Frequency
LLM	Large Language Model
MOS	Mean Opinion Score
NVML	NVIDIA Management Library

ONNX	Open Neural Network Exchange
PCM	Pulse-Code Modulation
RAM	Random Access Memory
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RSS	Resident Set Size
TTFA	Time-to-First-Audio
TTS	Text-to-Speech
VRAM	Video Random Access Memory
WER	Word Error Rate

Appendix A

The following is the primary system prompt used for the evaluation described in Section 3.6.

```
You are a concise, factual but friendly voice assistant. Everything you write is
converted to speech and read aloud, and~the user's words reach you through
automatic speech~recognition.

Stay on the topic of the user's request and within your assistant role at all times.
 If~the user asks for something unrelated to your role, steer back politely in one
short~sentence.

Start your reply with a very short sentence of at most five words that gives the
direct answer or acknowledges the request, and~end it with a full stop. If~that
fully answers the user, stop there. Otherwise add one or two short sentences of
detail. Never use more than three sentences in~total.

The user's words come from a speech recognizer and may contain a misheard word or
missing punctuation. Silently interpret obvious recognition errors from context. If
~the request is unintelligible or critically ambiguous, use your reply to ask one
brief clarifying question instead of guessing; confirming a word you are unsure of
is allowed. Never comment on transcription~quality.

Speak for the ear: short sentences, everyday words, no lists, no markdown, no
symbols, no emojis, and~no abbreviations. Write numbers and units as words where
natural. If~you are not sure of a fact, say so briefly instead of inventing~details.


Answer in English. Do not simulate a conversation or write lines for the user. Give
 a single reply only.
```

Appendix B. Implementation Settings and Preparatory Checks

This appendix specifies model tags, runtime settings, preparatory calibrations, and execution interfaces for the protocol in Section 3. Calibration measurements characterize component settings; they are distinct from comparisons on the 999-item scored evaluation set (Section 3.7). File paths refer to the archived code snapshot used for the reported runs (project repository in Data Availability). Command templates expose the required inputs and parameters without prescribing unverified experiment-specific paths or seeds.

Appendix B.1. Model Tags, Matrix, and Fixed Settings

Exact model tags combine the prefixes and suffixes in Table A1. The reported model-size/decoding comparison uses 24 Vosk cells and 12 Whisper cells. The preci-

sion comparison uses Llama 3.1 8 B, Qwen 2.5 7 B, and Gemma 3 4 B at Q4_K_M, Q8_0, and FP16, all at temperature 0. The repository lists 44 YAML files (28 Vosk-transcript and 16 Whisper-transcript), including greedy/sampled Vosk variants and extra Q8_0/FP16 files that are not a third factorial. The authoritative cell definition is its YAML, named in `configs/configs.csv`; descriptive CSV columns are not independently executed.

Table A1. Tag construction in the configuration matrix. Concatenate a prefix, a colon, and the listed suffix; braces indicate alternatives, not literal tag characters.

Prefix	Q4 Suffixes	Additional Suffixes
`llama3.2`	`1b-instruct-q4_K_M`; `3b-instruct-q4_K_M`	—
`llama3.1`	`8b-instruct-q4_K_M`	`8b-instruct-q8_0`; `8b-instruct-fp16`
`qwen2.5`	`{1.5b,3b,7b,32b}-instruct-q4_K_M`	`7b-instruct-q8_0`; `7b-instruct-fp16`
`gemma3`	`{1b,4b,27b}-it-q4_K_M`	`4b-it-q8_0`; `4b-it-fp16`

Shared settings of the configurations used for the reported runs are listed in Table A2.

Table A2. Shared settings of the configurations used for the reported runs. Inactive engine options can remain in a YAML without making that engine part of the run.

Setting	Value
Input and pacing	File input, real-time delivery, endpoint trigger
Audio chunks/Vosk silence	100 ms/flat 600 ms
Vosk asset	`vosk-model-small-en-us-0.15`
Whisper options	`small`, CUDA, explicit `int8`
Piper voice	`en_US-lessac-medium.onnx`, matching JSON sidecar
Piper execution	CPU, 12 intra-operation threads; playback off
LLM token/context limits	150/1024
LLM temperature/seed	Greedy: 0, unset seed; sampled: 0.7, seed 42
Model retention	`30 m`
Offload/batch/runner threads	Unset (server defaults); model measured 100% GPU-resident in every launch
Maximum text chunk	140 characters
Prompt file	`prompts/short-opener.txt`; text in Appendix A

The prompt file and external model/audio assets are not all distributed in the tracked source tree. Appendix A supplies the prompt body. Archived `system_prompt.txt` files prepend a source-label line before that body; that label is not part of the instruction to recreate. Audio, Vosk, Piper, and Ollama model files must be provisioned before launching the examples. Tags should be accompanied by the resolved model identity in the experiment archive.

Appendix B.2. Endpoint and Audio Preparation

The preparatory Vosk sweep used 833 read-aloud recordings and a confirmed subset of 577, where a non-endpointed Vosk transcript agreed with an independent faster-whisper transcript and covered the detected speech. It compared non-empty final-result splits with a separate VAD speech-end reference. Conditioning on this agreement supports endpointer tuning but also limits generalization to recognition failures and spontaneous speech. The calibration recordings are HeySQuAD recordings of questions from SQuAD training articles, whereas the evaluation corpus comes from the validation split (Section 3.7); the two sets therefore share no questions or articles. In the sweep, the endpoint delay d of a recording runs from the end of the last recognized word of the closing final result to the audio position at which that result is published. It is measured offline in audio time,

unlike the VAD-anchored, wall-clock `stt_endpoint_delay` of the evaluation runs; d_{50} and d_{99} denote its median and 99th percentile over the confirmed recordings.

With Vosk 0.3.44, the implementation applies a nonzero silence setting through a temporary model configuration because the setting is read at model initialization. A single number assigns the same silence duration to Kaldi rules 2–4; three slash-separated values preserve a staged schedule. Zero leaves the model configuration unchanged. Rule confidence conditions make a staged schedule different from a flat minimum-pause threshold.

Table A3 reports the preparatory endpointer calibration on the confirmed recordings.

Table A3. Preparatory endpointer calibration on 577 confirmed recordings at 100 ms chunks. Delays are percentiles of the recognizer-anchored endpoint delay d, not model-comparison latency estimates.

Schedule (ms)	Split Rate	p50 (ms)	p90 (ms)	p99 (ms)
Model defaults: 500/750/1000	0.52%	990	1130	1200
Flat 600	0.17%	930	1030	1080
500/600/600	0.52%	880	1010	1060
500/550/600	0.69%	850	960	1020

The flat 600 ms setting was retained as a delay–split compromise, not as a universal optimum. The faster gated settings reduce the median delay but increase premature splits in this calibration. One false-start recording remains split even at longer thresholds. Consequently, neither increasing the silence threshold nor removing difficult recordings can establish correct turn-taking for unrestricted users.

Table A4 repeats the flat-threshold sweep offline on the evaluation recordings themselves, at the evaluation chunk length of 100 ms and thresholds of 300–800 ms. At every threshold the endpointer fired before the end of the file on all 999 recordings (minimum margin of 182 ms at 800 ms); higher thresholds would need more trailing silence than the evaluation files carry and are not reported. At 600 ms the offline firings reproduce the real-time ASR pass: measured from the same VAD speech end, the real-time `stt_endpoint_delay` exceeds the offline delay by a median of 20 ms (5th–95th percentile 15–44 ms; Spearman $\rho = 0.94$), and both split the same four recordings. The word error rate against the intended question is 0.201–0.202 at every threshold, against 0.200 in the real-time pass (Section 4.1).

Table A4. Offline sweep of the flat silence threshold on the evaluation recordings (100 ms chunks; 999 recordings). Confirmed recordings (884) are defined as in the calibration sweep: the non-endpointed Vosk transcript agrees with faster-whisper and covers the detected speech. Split mid-sentence: share of the confirmed recordings on which the endpointer fired before the VAD speech end. Delays are medians over the confirmed recordings, measured from the last recognized word (d_{50}) and from the VAD speech end.

	Flat Silence Threshold (ms)					
	300	400	500	600	700	800
Split mid-sentence	4.86%	1.70%	0.68%	0.34%	0.11%	0.11%
d_{50} (ms)	640	720	810	940	1020	1120
Median from VAD end (ms)	656	736	832	952	1032	1136

Appendix B.2.1. Chunk Length and Empirical Delay Relation

Least-squares fits over 32 combinations of six chunk lengths (50–250 ms) and flat thresholds of 200–1500 ms give

$$d_{50} \approx s + 280\,\text{ms} + 0.40c, \tag{A1}$$

$$d_{99} \approx s + 390\,\text{ms} + 0.75c, \tag{A2}$$

which reproduce the measured percentiles to within about 35 ms and the d_{50} row of Table A4 to within 20 ms. Here s is the flat silence setting and c the chunk duration, both in milliseconds. The constant term is the recognizer's lag between the end of the last word and the trailing silence that its endpoint rule scores; the chunk term arises because an endpoint can be detected only at a chunk boundary. These empirical fits summarize the calibration range; they are not physical constants or guarantees for a different corpus. A second offline sweep on the evaluation recordings varied the chunk length at the 600 ms threshold. Chunks of 50, 100, 150, 200, and 250 ms gave d_{50} of 920, 940, 960, 970, and 1010 ms (0.42 ms per millisecond of chunk, against 0.40 in Equation (A1)), with the same three premature firings among the 884 confirmed recordings and a word error rate against the intended question of 0.200–0.201 at every length, so chunk length acts as a latency setting only. Its 100 ms pass repeats the 600 ms setting of Table A4 and reproduces 994 of the 999 firing times exactly. The 100 ms chunk choice limits boundary quantization without selecting the smallest tested chunk solely for a modest delay improvement.

Appendix B.2.2. Speech-End Anchor

TTFA and endpoint delay share the VAD speech-end anchor, so moving it by Δ changes both by Δ and leaves TTFC and first-chunk synthesis unchanged. On the validation subset a shift of ± 100 ms changes the endpoint share of TTFA by at most three percentage points (45–50% for `gemma3:27b`). Where Vosk follows the speech, its last recognized word is a close substitute: on the confirmed evaluation recordings the median delay from the VAD speech end exceeds d_{50} by only 12–22 ms (Table A4). Where recognition fails, an engine-derived boundary can move much further (Section 3.8.1), so the shared VAD anchor is retained.

Appendix B.2.3. Trailing Silence

At $s = 600$ ms and $c = 100$ ms, Equation (A2) gives a preparatory trailing-silence target of 1065 ms. The metadata validator compares the available tail with this requirement. Insufficient silence can force file-finalization behaviour that does not represent an endpoint occurring during live capture. Conversely, excess trailing silence still delays the buffered Whisper path and an explicit end-of-file trigger. It is therefore not generally free padding.

Each evaluation file carries about 1.5 s of appended digital silence, and every total tail exceeds this target (Section 3.7). In the real-time ASR pass that produced the frozen Vosk transcripts, the endpoint fired before the end of the file on all 999 scored recordings (minimum margin of 354 ms). For 877 recordings (87.8%) it fired inside the appended silence. Even so, the median endpoint delay was about 960 ms for the 54 recordings whose natural trailing silence alone exceeded 1.2 s and about 980 ms for the 244 with less than 0.3 s; the digital padding did not shorten the measured endpoint. Those truncated files remain excluded; valid CSV timing fields alone do not reinstate them.

Appendix B.3. Corpus-Screening Details

The answerability-filter report documents 4158 $\rightarrow$ 1002 using only the exact `False` value of `is_impossible`. The subsequent question-type report tokenizes the intended question case-insensitively, normalizes regular interrogative contractions, merges *who/whom/whose*, and assigns a broad type using the first exact interrogative form. It classifies the 1002 records as: *what* 638, *how* 114, *who* 84, *when* 71, *which* 37, *where* 29, *why* 19,

yes/no 4, and other 6. These are pre-mismatch counts and are not quotas used to construct the final corpus.

The mismatch report independently transcribes every WAV with faster-whisper `small.en`. It maps content tokens to standard American Soundex codes, first collapsing consecutive one-letter tokens representing spelled acronyms. A record is a candidate if the independent transcript shares strictly more distinct codes with the source `transcription` than with the intended `question`. Exclusion additionally requires the two overlap-evidence sets to be exactly disjoint. Ties, intersecting evidence, and unresolved cases remain.

Within the mismatch-checking procedure, the automatic rule flags 44 candidates and excludes one clear content mismatch, `5729fc3d1d046914007796a0.wav`, giving 1002 → 1001. The remaining flagged cases are retained because the observed differences are limited to minor transcription variations, such as acronym spelling, named-entity recognition, identifier formatting, or isolated ASR substitutions, while the intended question content remains unchanged. During the subsequent evaluation runs, two additional defective recordings, `572a0bfaaf94a219006aa77a.wav` and `5729e500af94a219006aa6b5.wav`, were identified.

The selection history is documented in the repository by two accompanying README files: a corpus README listing all WAV files included in the corpus, and a filtering README describing the reduction from 4158 source records to 1002 answerable items, the pre-run mismatch reduction to 1001, and the subsequent identification of two defective recordings during evaluation runs. Protocol settings and repetition counts are documented separately in Section 3.11.

Appendix B.4. Synthesis and Serving Preparation

Appendix B.4.1. Piper Thread Count

The preparatory thread-count calibration compares the default ONNX Runtime pool with explicit thread counts while `phi3:mini` is streaming, using ten alternating passes of nine repetitions and one process per condition. First-chunk synthesis averages are 226 ms with the default pinned pool, 170 ms with 16 explicit threads, and 123 ms with 12. The corresponding idle values are 121, 120, and 127 ms. The contrast motivates testing under concurrent load: an idle microbenchmark does not reveal the same contention. The selected 12-thread setting is specific to the documented host and workload, not a universal scaling rule.

Appendix B.4.2. Package and Device Setup

The CPU requirements pin Vosk 0.3.44, faster-whisper 1.2.1, Piper 1.4.2, NumPy 2.2.6, SciPy 1.15.3, SoundFile 0.14.0, psutil 7.2.2, Requests 2.34.2, PyYAML 6.0.3, sounddevice 0.5.1, and nvidia-ml-py 13.610.43. These are repository environment specifications, not an automatically captured installed-package inventory. `ffmpeg` is an external dependency. A CPU or CUDA provider must actually initialize on the host architecture; nominal CUDA availability alone is insufficient to establish execution of the requested kernels. The present matrix avoids making alternate TTS providers an experimental factor.

Appendix B.5. Resource-Field Interpretation

Table A5 lists the resource columns and their measurement scope.

The CPU/GPU windows align with the timed worker spans for `stt`, `llm_ttft`, `llm_first_chunk_fill`, `tts_first_chunk`, and `e2e_response_ready`. Other rows reuse nearby observations: `llm_ttfc` uses the first-chunk-fill window; `ttfa` reuses first-chunk synthesis; `stt_endpoint_delay` reuses recognition; `llm_prompt_eval` reuses the first-token

snapshot; complete generation and synthesis rows use worker-end snapshots. Their resource fields therefore do not all describe an interval equal to their `duration_ms`.

Table A5. Resource columns and their measurement scope. Memory columns are snapshots, not peaks; blank values indicate unavailable observations.

Field	Scope
`cpu_percent`	System CPU rate over the associated sampling window
`ram_percent`	System memory utilization at the snapshot
`rss_mb`	Python pipeline resident set
`llm_rss_mb`	Sum of discoverable Ollama server and runner RSS
`llm_vram_mb`	GPU memory attributed to those processes
`llm_model_vram_mb`	Model residency reported by `/api/ps`
`gpu_util_percent`	Sampled GPU utilization over the associated window
`gpu_mem_used_mb`	Device-allocated memory from NVML v2 when available
`gpu_mem_total_mb`	Device memory capacity

NVML sampling runs at 0.1 s intervals with available process or device utilization information. The driver's own sampling resolution remains a limit, especially for short stages. The fallback queries `nvidia-smi` once per second outside the timed path. These mechanisms do not log complete thermal or frequency trajectories and do not support a claim that such variables were experimentally held constant.

Appendix B.6. Execution Interfaces and Provenance

The following commands are templates for the implemented interfaces. Replace `CORPUS` with the prepared directory containing only eligible files and matching metadata, `ASR_RUN` with the selected timestamped ASR output, and `MATRIX_CSV` with the matrix for the intended comparison set. Use three rounds for the model-size/decoding comparison and record the selected `SEED`. The parameters `N` and `R_VAL` denote the independently specified item and repetition counts for TTFA validation. Provision the prompt and model assets before execution. A dry run checks scheduling, not inference or completeness.

```
# Realtime ASR-only arm, using one of the evaluated configurations.
python assistant.py --config configs/01-llama3_2_1b_instruct_q4_K_M-t0_0-vosk_cpu.yaml --audio CORPUS --
      asr-only --out-dir outputs/asr

# Preview the repeated-cell schedule.
python run_campaign.py --configs-csv MATRIX_CSV --audio-dir CORPUS --out-dir outputs/matrix --rounds 3 --
      seed SEED~--dry-run

# Downstream evaluation: use recognizer-matched cells and their frozen input.
python run_campaign.py --cells 01-llama3_2_1b_instruct_q4_K_M-t0_0-vosk_cpu --audio-dir CORPUS --out-dir
      outputs/text-vosk --rounds 3 --seed SEED -- --input-mode text --transcripts ASR_RUN/transcripts.
      jsonl

# Join one canonical ASR arm with downstream timing.
python ttfa_reconstruct.py --asr-arm outputs/asr --canonical ASR_RUN --text-arm outputs/text-vosk --offset
      -ms~5

# Preview a split-versus-full validation subset and its launches.
python ttfa_validation.py --audio-dir CORPUS --out-dir outputs/validation --cells 01-
      llama3_2_1b_instruct_q4_K_M-t0_0-vosk_cpu --items 200 --replicates 3 --seed 42 --sampling random~--
      dry-run

# Offline endpointer sweeps over flat thresholds and chunk lengths on the evaluation recordings.
python vosk_endpoint_sweep.py --audio-dir CORPUS --vosk-model vosk/vosk-model-small-en-us-0.15 --chunk-ms
      100 --t-end-ms 300 400 500 600 700 800 --out-dir outputs/endpoint_sweep
python vosk_endpoint_sweep.py --audio-dir CORPUS --vosk-model vosk/vosk-model-small-en-us-0.15 --chunk-ms
      100 50 150 200 250 --t-end-ms 600 --out-dir outputs/endpoint_sweep_chunk
```

A Vosk transcript arm and a Whisper transcript arm use different canonical source files and matching cell subsets. Supplying one Vosk transcript file to all matrix cells would not create a recognizer comparison, even if some YAML filenames contain `whisper`. The recorded transcript provenance, rather than the filename alone, determines the actual input arm.

A run manifest specifies the selected cell set, its round count, and execution order under Ollama 0.33.3. The model-size/decoding and weight-precision comparisons retain separate analysis mappings even where their model choices overlap. The TTFA-validation experiment likewise records its own repetitions rather than inheriting a launch count from another analysis.

Outputs are organized as `<output_dir>/<cell>/r<round>/<timestamp>/`, with resolved configurations, stage CSVs, replies, and retained waveforms per launch. Cell and launch identifiers travel on log rows. Resumption skips launches recognized as complete; that operational check is not proof of item-complete stages. The analysis archive keeps the configuration matrix, resolved YAMLs, prompts, canonical ASR transcripts, metadata, the exclusion list, the run manifest, and raw logs together.

The reconstruction script joins by recording identity after removing each run's first item from the timed analysis. When several ASR runs are available, one must be selected explicitly as canonical; their endpoint delays are not silently averaged into a new input arm. Additional ASR runs can describe endpoint variability. Duplicate copies of identical CSV inputs are detected, and incomplete matches are reported rather than replaced by zero. Filename-based joins do not independently verify transcript content, so the downstream arm must use the exact frozen transcript artefact associated with its canonical ASR run.

Implementation Map

`assistant.py` defines mode dispatch, provenance, timing, and cancellation; `audio_sources.py` and `stt_engines.py` define audio delivery and recognition; `llm_engine.py`, `pipeline_channels.py`, and `tts_engines.py` define generation, worker communication, and synthesis. `run_campaign.py` schedules independent launches, and `campaign_report.py` summarizes their variation. `ttfa_residual_check.py`, `ttfa_reconstruct.py`, and `ttfa_validation.py` implement the within-run, reconstructed, and split-versus-full checks, respectively. Audio preparation is supported by `make_metadata.py`, `validate_metadata.py`, and `vosk_endpoint_sweep.py`. These are the sources for the architecture figure, setting tables, and command examples rather than illustrative code for unimplemented experiments.

References

1. Mu, B.; Guo, P.; Sun, Z.; Wang, S.; Liu, H.; Shao, M.; Xie, L.; Chng, E.S.; Xiao, L.; Feng, Q.; et al. Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods. In *Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2026; pp. 19442–19446.
2. European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data. *Off. J. Eur. Union* **2016**, *L 119*, 1–88.
3. Maslych, M.; Katebi, M.; Lee, C.; Hmaiti, Y.; Ghasemaghaei, A.; Pumarada, C.; Palmer, J.; Segarra Martinez, E.; Emporio, M.; Snipes, W.; et al. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*; Association for Computing Machinery: New York, NY, USA, 2025. [CrossRef]
4. Peng, Z.; Mo, K.; Zhu, X.; Chen, J.; Chen, Z.; Xu, Q.; Ma, X. Understanding User Perceptions of Robot's Delay, Voice Quality-Speed Trade-off and GUI during Conversation. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–8. [CrossRef]

5. Boukaram, H.A.; Ziadee, M.; Sakr, M.F. Mitigating the Effects of Delayed Virtual Agent Response Time Using Conversational Fillers. In *Proceedings of the 9th International Conference on Human-Agent Interaction*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 130–138. [CrossRef]
6. He, Y.; Sainath, T.N.; Prabhavalkar, R.; McGraw, I.; Alvarez, R.; Zhao, D.; Rybach, D.; Kannan, A.; Wu, Y.; Pang, R.; et al. Streaming End-to-End Speech Recognition for Mobile Devices. In *Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, NY, USA, 2019; pp. 6381–6385. [CrossRef]
7. Shangguan, Y.; Prabhavalkar, R.; Su, H.; Mahadeokar, J.; Shi, Y.; Zhou, J.; Wu, C.; Le, D.; Kalinli, O.; Fuegen, C.; et al. Dissecting User-Perceived Latency of On-Device E2E Speech Recognition. In *Proceedings of the Interspeech 2021*; International Speech Communication Association (ISCA): Grenoble, France, 2021; pp. 4553–4557. [CrossRef]
8. Bijwadia, S.; Chang, S.Y.; Li, B.; Sainath, T.N.; Zhang, C.; He, Y. Unified End-to-End Speech Recognition and Endpointing for Fast and Efficient Speech Systems. In *Proceedings of the 2022 IEEE Spoken Language Technology Workshop*; IEEE: New York, NY, USA, 2022; pp. 310–316. [CrossRef]
9. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 28492–28518.
10. Wang, R.; Xu, Z.; Lin, F.X. WhisperFlow: Speech Foundation Models in Real Time. *arXiv* **2025**. [CrossRef]
11. Alpha Cephei. Vosk Offline Speech Recognition API. 2024. Available online: https://alphacephei.com/vosk/ (accessed on 10 August 2026).
12. Kim, J.; Kong, J.; Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 5530–5540.
13. Huang, F.; Chen, D.; Chen, C.; Song, Z.; Lin, G.; Zeng, K. EmbSpeech: A Unified Framework Towards Low-Resource Zero-Shot Speech Synthesis. *Adv. Intell. Comput. Technol. Appl.* **2025**, *15864*, 173–184. [CrossRef]
14. International Telecommunication Union. *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*; Recommendation P.85; ITU-T: Geneva, Switzerland, 1994.
15. International Telecommunication Union. *Subjective Evaluation of Speech Quality with a Crowdsourcing Approach*; Recommendation P.808; ITU-T: Geneva, Switzerland, 2021.
16. Hansen, M. Piper: A Fast, Local Neural Text to Speech System. 2023. Available online: https://github.com/rhasspy/piper (accessed on 10 August 2026).
17. Zhou, Z.; Ning, X.; Hong, K.; Fu, T.; Xu, J.; Li, S.; Lou, Y.; Wang, L.; Yuan, Z.; Li, X.; et al. A Survey on Efficient Inference for Large Language Models. *arXiv* **2024**. [CrossRef]
18. Ollama. Ollama. 2024. Available online: https://github.com/ollama/ollama (accessed on 10 August 2026).
19. Gerganov, G. llama.cpp. 2023. Available online: https://github.com/ggml-org/llama.cpp (accessed on 10 August 2026).
20. Gong, R.; Ding, Y.; Wang, Z.; Lv, C.; Zheng, X.; Du, J.; Qin, H.; Guo, J.; Magno, M.; Liu, X. A Survey of Low-bit Large Language Models: Basics, Systems, and Algorithms. **arXiv** *2025*. [CrossRef]
21. Jin, R.; Du, J.; Huang, W.; Liu, W.; Luan, J.; Wang, B.; Xiong, D. A Comprehensive Evaluation of Quantization Strategies for Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand and Virtual Meeting*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 12186–12215. [CrossRef]
22. Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; Hou, L. Instruction-Following Evaluation for Large Language Models. *arXiv* **2023**. [CrossRef]
23. Jiang, Y.; Wang, Y.; Zeng, X.; Zhong, W.; Li, L.; Mi, F.; Shang, L.; Jiang, X.; Liu, Q.; Wang, W. FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 4667–4688. [CrossRef]
24. Wang, Y.Y.; Acero, A.; Chelba, C. Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy? In *Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding*; IEEE: New York, NY, USA, 2003; pp. 577–582. [CrossRef]
25. Lee, C.H.; Wu, S.L.; Liu, C.L.; Lee, H.y. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3459–3463. [CrossRef]
26. You, C.; Chen, N.; Liu, F.; Ge, S.; Wu, X.; Zou, Y. End-to-end Spoken Conversational Question Answering: Task, Dataset and Model. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 July 2022; pp. 1219–1232. [CrossRef]
27. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Kerrville, TX, USA, 2023. [CrossRef]

28. Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; Zeghidour, N. Moshi: A Speech-Text Foundation Model for Real-Time Dialogue. *arXiv* **2024**, arXiv:2410.00037. [CrossRef]
29. Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; Feng, Y. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. *arXiv* **2024**, arXiv:2409.06666. [CrossRef]
30. Xie, Z.; Wu, C. Mini-Omni: Language Models Can Hear, Talk While Thinking in Streaming. *arXiv* **2024**, arXiv:2408.16725. [CrossRef]
31. Héder, M.; Rigó, E.; Medgyesi, D.; Lovas, R.; Tenczer, S.; Török, F.; Farkas, A.; Emődi, M.; Kadlecsik, J.; Mező, G.; et al. The Past, Present and Future of the ELKH Cloud. *Inf. Társad.* **2022**, *22*, 128. [CrossRef]
32. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*; IEEE Signal Processing Society: New York, NY, USA, 2011.
33. Guillaume, K. faster-Whisper: Faster Whisper Transcription with CTranslate2. 2023. Available online: https://github.com/SYSTRAN/faster-whisper (accessed on 10 August 2026).
34. OpenNMT. CTranslate2: Fast Inference Engine for Transformer Models. Software. Available online: https://github.com/OpenNMT/CTranslate2 (accessed on 16 August 2026).
35. Yuan, J.; Li, H.; Ding, X.; Xie, W.; Li, Y.J.; Zhao, W.; Wan, K.; Shi, J.; Hu, X.; Liu, Z. Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 169819–169851.
36. Ouyang, S.; Zhang, J.M.; Harman, M.; Wang, M. An empirical study of the non-determinism of chatgpt in code generation. *Acm Trans. Softw. Eng. Methodol.* **2025**, *34*, 1–28. [CrossRef]
37. Wu, Y.; Rallabandi, S.; Srinivasamurthy, R.; Dakle, P.P.; Gon, A.; Raghavan, P. HeySQuAD: A Spoken Question Answering Dataset. *arXiv* **2023**. [CrossRef]
38. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2016; pp. 2383–2392. [CrossRef]
39. Flesch, R. A New Readability Yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef] [PubMed]
40. Kincaid, J.P.; Fishburne, R.P.; Rogers, R.L.; Chissom, B.S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*; Technical Report Research Branch Report 8-75; Naval Technical Training Command: Millington, TN, USA, 1975.
41. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*; Association for Computational Linguistics: Kerrville, TX, USA, 2004; pp. 74–81.
42. Liu, C.W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; Pineau, J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2016; pp. 2122–2132. [CrossRef]
43. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
44. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
45. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
46. Cliff, N. Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychol. Bull.* **1993**, *114*, 494–509. [CrossRef]
47. Lakens, D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc. Psychol. Personal. Sci.* **2017**, *8*, 355–362. [CrossRef] [PubMed]
48. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: New York, NY, USA, 1993.
49. Dean, J.; Barroso, L.A. The Tail at Scale. *Commun. ACM* **2013**, *56*, 74–80. [CrossRef]
50. Beyer, B.; Jones, C.; Petoff, J.; Murphy, N.R. (Eds.) *Site Reliability Engineering: How Google Runs Production Systems*; O'Reilly Media: Santa Rosa, CA, USA, 2016.
51. Hutson, A.D. Calculating Nonparametric Confidence Intervals for Quantiles Using Fractional Order Statistics. *J. Appl. Stat.* **1999**, *26*, 343–353. [CrossRef]
52. Meta AI. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. 2024. Available online: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (accessed on 15 August 2026).
53. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**. [CrossRef]
54. Qwen Team. Qwen2.5 Technical Report. *arXiv* **2025**. [CrossRef]
55. Gemma Team. Gemma 3 Technical Report. *arXiv* **2025**. [CrossRef]
56. Shrout, P.E.; Fleiss, J.L. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychol. Bull.* **1979**, *86*, 420–428. [CrossRef] [PubMed]

57. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
58. Bland, J.M.; Altman, D.G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *327*, 307–310. [CrossRef]
59. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef] [PubMed]
60. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM Reporting Guideline for Studies Using Large Language Models. *Nat. Med.* **2025**, *31*, 60–69. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.