CONVOAUDIT: A CONTROLLED DIAGNOSTIC BENCHMARK FOR CONVERSATIONAL AUDIO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding conversational audio is essential for intelligent systems to support effective communication and assistance. However, evaluating this capability requires assessing not only the content they recover but also how they track participants, interpret interactions, and locate events in time. Although existing benchmarks offer increasingly detailed assessments, aggregate scores can obscure selective weaknesses, while multiple factors always vary together in real-world conversational audio, making their individual effects difficult to isolate. In this work, we introduce ConvoAudit, a controlled diagnostic benchmark for conversational audio understanding, to address these challenges. Specifically, ConvoAudit evaluates dialogue content, interaction structure, and acoustic-temporal organization as three complementary dimensions to profile system capabilities under different conversational conditions. These conditions are realized via synthesized variants of shared source dialogues, thus enabling us to examine how capability profiles and performance gaps between transcription-based and direct-audio systems vary with conversational organization. Our evaluation on over fifty system configurations reveals that transcription-based and direct-audio systems can achieve similar scores on dialogue content yet differ substantially on interaction structure and acoustic–temporal organization, with the relative advantage of direct-audio systems varying across conversational conditions. Matched transcript controls further demonstrate that the usefulness of structural evidence depends on both its availability and its correctness. Fine-tuning on ConvoAudit's training dataset also enables open-weight models to outperform strong proprietary baselines, highlighting its value for targeted model improvement. Together, these analyses connect where systems struggle with controlled tests of which evidence improves their judgments, thereby supporting more informative evaluation of conversational audio understanding.

1 INTRODUCTION

As speech technologies expand in various real-world applications (e.g., customer service, meeting assistance, and voice interaction), conversational audio understanding is becoming the foundational capability for intelligent systems (Chu et al., 2023; Ding et al., 2025). Recent advances have enabled diverse approaches to developing this capability, including transcription-based systems that combine automatic speech recognition (ASR) (Malik et al., 2021) with large language models (LLMs) (Zhao et al., 2026), as well as direct-audio systems based on large audio-language models (LALMs) (Yang et al., 2025). Alongside this progress, accurate evaluation has become increasingly important, as it helps establish the capabilities of existing systems, identify the conditions under which they struggle, and guide further development (Peng et al., 2025).

Several representative benchmarks have been introduced to evaluate audio understanding in recent years. Specifically, MMAU (Sakshi et al., 2025) provides broad coverage of perception and reasoning across speech, sound, and music. MMAR (Ma et al., 2026) focuses on multi-step reasoning over these domains and their mixtures, while MMSU (Wang et al., 2026) assesses fine-grained perception and reasoning through linguistically grounded tasks. However, existing benchmarks predominantly rely on real-world audio recordings and often summarize system performance using aggregate scores. Despite their advances in coverage and task granularity, an important diagnostic question remains:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

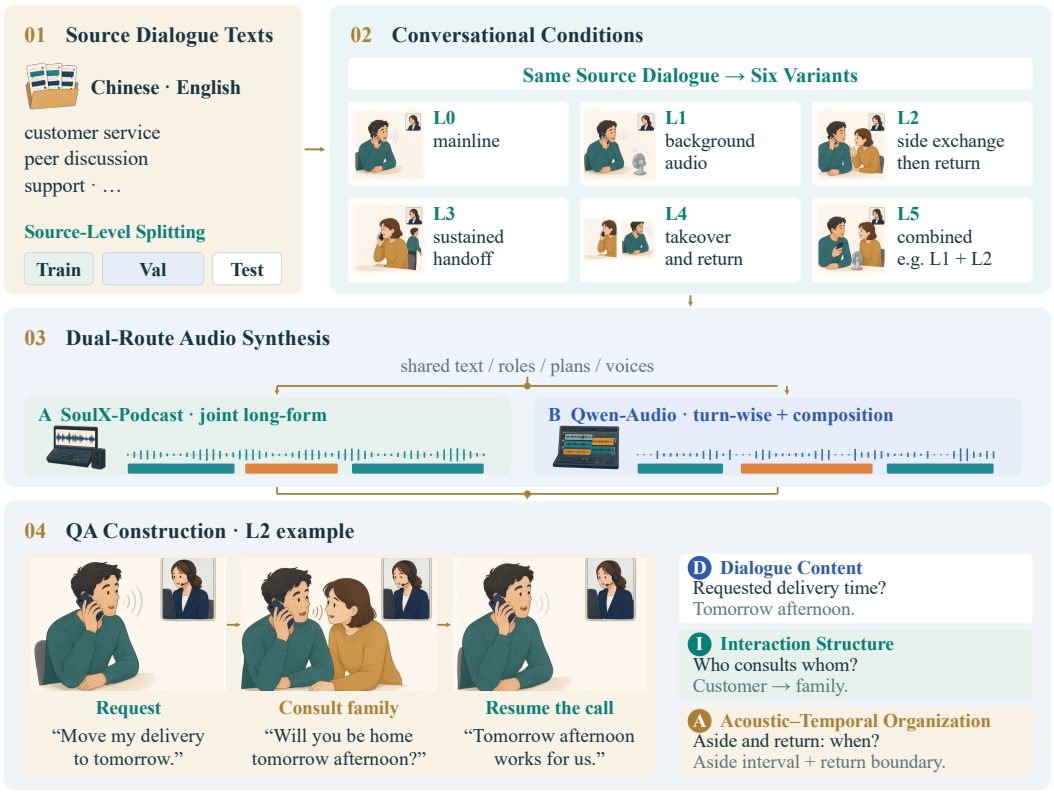


Figure 1: The comprehensive pipeline for constructing the ConvoAudit benchmark. Curated source dialogue texts are expanded into six conversational variants with the shared mainline, then realized through two audio-synthesis routes to support cross-implementation consistency checks. The resulting audio is paired with questions covering dialogue content (D), interaction structure (I), and acoustic–temporal organization (A), enabling capability profiling across conversational conditions. Specifically, an L2 example in customer-service scenario is used here to illustrate the QA construction stage.

how do systems' capability profiles vary with conversational organization? For example, the same customer-service request may involve consulting a family member, handing over the phone, or returning after a temporary takeover. The request remains unchanged, but participation, continuity, and event timing differ. Aggregate scores can conceal weaknesses specific to these conditions, while comparisons across heterogeneous real-world recordings can confound conversational organization with content and acoustics (Luo et al., 2026; Gosai et al., 2026).

To address these challenges, we introduce **ConvoAudit**, a controlled diagnostic benchmark connecting *capability profiles*, *conversational conditions*, and *evidence interventions* for conversational audio understanding. As presented in Figure 1, we collect and curate source dialogue texts across diverse conversational settings, then construct six variants of each source that preserve its mainline content while varying conversational organization. To support reliable system comparisons, we realize shared conversational plans through two audio synthesis routes, enabling consistency checks across synthesis implementations. Based on the synthesized conversation audio, we further construct question–answer pairs that assess three complementary dimensions: *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A). The resulting capability profiles enable comparisons of various systems across shared-source conditions, moving beyond aggregate scores to reveal *where systems struggle and how these weaknesses vary with conversational organization*.

Based on the ConvoAudit benchmark, we evaluate over 50 system configurations, spanning proprietary and open-weight direct-audio models, ASR–LLM cascades, and benchmark-adapted variants. Our results show that transcription-based and direct-audio systems can achieve similar dialogue-content scores yet differ substantially in interaction structure and acoustic–temporal organization. Moreover, the relative advantage of direct-audio systems varies across conversational conditions, with temporary

takeovers and returns posing substantial challenges for both pathways. Matched transcript controls further show that the value of structural evidence depends on both its availability and its correctness: timestamps support temporal localization, while the usefulness of speaker information depends on the accuracy of speaker assignments. We also find that open-weight models fine-tuned on ConvoAudit's training dataset can outperform strong proprietary baselines, highlighting its value for both diagnostic evaluation and targeted model improvement. Together, these findings connect condition-dependent capability gaps with controlled tests of which evidence improves system judgments, enabling more informative evaluation of conversational audio understanding.

In summary, our main contributions are as follows:

- We introduce ConvoAudit, a controlled diagnostic benchmark for conversational audio understanding that supports systematic analysis of capability profiles and their variation across conversational conditions.
- We evaluate over 50 system configurations, revealing that similar dialogue-content scores can conceal substantial differences in interaction structure and acoustic–temporal organization, while the relative advantage of direct-audio systems varies across conversational conditions.
- We show that open-weight models fine-tuned on ConvoAudit's training dataset outperform strong proprietary baselines, thereby highlighting its value not only for diagnostic evaluation but also for targeted model improvement.

2 Related Work

2.1 Conversational Audio Understanding

Conversational audio understanding is increasingly important for intelligent systems to interpret user needs and respond appropriately across real-world applications, including customer service (Gung et al., 2023), meeting assistance (Thonet et al., 2025), and voice interaction (Zhang et al., 2023). To support these applications, various approaches have been developed over the past few years. Among them, transcription-based systems are widely used, where the large language model (LLM) reasons over transcripts recovered by the automatic speech recognition (ASR) model such as Whisper (Radford et al., 2023) and SenseVoice (An et al., 2024), optionally enriched with word-level alignment from tools like WhisperX (Bain et al., 2023). More recently, advances in large audio-language models (LALMs) have enabled direct-audio systems that reason over encoded audio rather than relying solely on ASR transcripts (Chu et al., 2023; Gong et al., 2024). Besides, various omni-modal models, such as Qwen3-Omni (Xu et al., 2025) and Baichuan-Omni-1.5 (Li et al., 2025), further integrate the perception of text, images, audio, and video with streaming text and speech generation, extending these capabilities to broader multimodal interactions. However, conversational audio is inherently complex, with variations in speaker participation, interaction patterns, and event timing placing different demands on system capabilities. Despite these advances, accurately assessing how well these systems understand remains an important challenge (Gao et al., 2025).

2.2 Audio Understanding Benchmarks

Reliable benchmarks are essential for assessing progress in audio understanding and guiding model development. Early efforts, such as AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2025), provide broad assessments spanning audio comprehension and speech, scene, and voice understanding. More recently, MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2026), and MMSU (Wang et al., 2026) further emphasize cross-domain perception and reasoning, multi-step reasoning, and fine-grained linguistic and paralinguistic skills, respectively. Complementing these broad assessments, MSUBench (Sun et al., 2026) focuses on speaker-centric understanding, TAG-Bench (Dai et al., 2026) evaluates temporal audio grounding, and Audio MultiChallenge (Gosai et al., 2026) assesses performance in natural multi-turn interactions. However, many existing benchmarks rely on real-world audio recordings and use aggregate scores to summarize system performance. Such scores can obscure capability-specific weaknesses, while differences in content and acoustics across real-world audio recordings also complicate controlled analysis of how system capability profiles vary with conversational organization. This highlights the critical need for controlled diagnostic benchmarks that isolate conversational organization from extraneous variations for fine-grained evaluations.

3 THE CONVOAUDIT BENCHMARK

3.1 OVERVIEW

ConvoAudit is a controlled diagnostic benchmark designed to reveal system strengths and weaknesses across various complementary capabilities and conversational conditions. It assesses three dimensions: *dialogue content* (D), covering facts, intentions, and discourse meaning; *interaction structure* (I), covering participating voices, channel relations, and continuity; and *acoustic–temporal organization* (A), covering voice and event attributes, the audible scene, and event timing.

To examine how these capabilities vary with conversational organization, ConvoAudit pairs shared-source dialogue variants with dimension-specific questions. Six conversational conditions preserve each source's substantive mainline while varying background activity, side exchanges, and speaker transitions. Two synthesis routes realize shared conversational plans, enabling consistency checks across implementations. This design supports comparisons of transcription-based and direct-audio systems under matched source conditions, revealing differences that the aggregate score can obscure.

3.2 BENCHMARK CONSTRUCTION

As presented in Figure 1, ConvoAudit is constructed through the following four stages:

1) Source-dialogue curation. We curate multi-turn dialogue texts in Chinese and English spanning diverse conversational settings, including customer service, peer discussions, and support interactions. Then we combine rule-based checks with model-assisted refinement to standardize utterances, correct turn boundaries, and assess semantic coherence. The curated dialogues provide diverse conversational mainlines that serve as the shared basis for constructing condition-specific variants. More information about source-dialogue curation can be found in Appendix A.1.

2) Conversational variant construction. For each source dialogue, we construct six conversational conditions while preserving its substantive mainline. In particular, L0 retains the main dialogue and its native paralinguistic events. L1–L4 respectively introduce noninteractive background audio, the side exchange followed by resumption, the persistent handoff, and the temporary takeover followed by return. L5 combines two to four of these components within the single dialogue, capturing more complex conversational scenarios in which multiple acoustic and interactional variations co-occur. More information about conversational variant construction can be found in Appendix A.2.

3) Dual-route audio synthesis. We use two complementary synthesis routes to examine whether the benchmark's diagnostic findings remain consistent across different acoustic realizations. In particular, Route-A jointly generates long-form audio with SoulX-Podcast-1.7B (Xie et al., 2025), whereas Route-B synthesizes individual turns with Qwen-Audio-3.0-TTS-Flash (Xiang et al., 2026) and composes them into the complete recording. Formally, for the source-dialogue text u_s, conversational condition $c \in \{\mathrm{L0}, \ldots, \mathrm{L5}\}$, and audio-synthesis route $r \in \{\mathrm{A}, \mathrm{B}\}$, we express the process as

$$p_{sc} = \mathcal{T}_c(u_s), \qquad (x_{scr}, z_{scr}) = \mathcal{R}_r(p_{sc}, v_s), \tag{1}$$

where $\mathcal{T}_c$ constructs the condition-specific conversational plan p_{sc}, and $\mathcal{R}_r$ realizes this plan as audio using synthesis route r. The resulting audio x_{scr} and associated realization records z_{scr} provide the basis for task references grounded in the voices, scenes, and events present in the recordings. Notably, conversational mainlines and event specifications remain matched across the two synthesis routes, enabling us to examine whether the observed capability patterns persist across acoustic realizations. More information about dual-route audio synthesis can be found in Appendix A.3.

4) Question–answer pair construction and quality checks. We construct question–answer pairs across the three capability dimensions using information derived from the preceding stages. In particular, we use constrained prompts to generate questions and reference answers grounded in the shared conversational mainline for *dialogue content* (D), allowing the resulting pairs to be reused across conversation variants. For *interaction structure* (I) and *acoustic–temporal organization* (A), we deterministically derive question–answer pairs from the voices, scenes, events, and timelines realized in each recording. We then apply task-specific quality checks to verify format compliance, supporting evidence, reference consistency under option reordering, and audio–timeline alignment. Besides, task eligibility is determined based on the availability of the required reference evidence, and construction metadata are excluded from standard model inputs when evaluating. More information about question–answer pair construction and quality checks can be found in Appendix A.4.

Benchmark Resources

Split	Recordings	Hours	QA items
Training	5,769	348.88	76,956
Validation	960	61.45	12,787
Test	1,200	74.30	16,185
Total	**7,929**	**484.63**	**105,928**

Test QA Distribution **16,185**

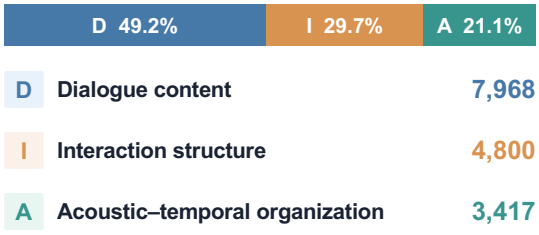


Figure 2: The statistics of our ConvoAudit benchmark. Left: recording counts, audio duration, and question–answer (QA) counts for the training, validation, and test sets. Right: distribution of test-set QA pairs across the three capability dimensions.

3.3 Benchmark Statistics

As summarized in Figure 2, ConvoAudit comprises 7,929 recordings totaling 484.63 hours and 105,928 question–answer instances. We partition the benchmark into training, validation, and test sets, containing 5,769, 960, and 1,200 recordings, respectively, with 76,956, 12,787, and 16,185 QA instances. All variants of the same source dialogue remain in the same split, as do detected lexical near-duplicate groups, thereby reducing cross-split overlap.

In particular, the test set accounts for 74.30 hours of audio, with its QA instances distributed across *dialogue content* (49.2%), *interaction structure* (29.7%), and *acoustic–temporal organization* (21.1%). Besides, it cover all six conversational conditions and both synthesis routes for each included dialogue, thereby supporting matched comparisons across conversational conditions and acoustic realizations. For rapid and low-cost preliminary evaluation, we also provide the test-mini set, a subset of 3,237 QA instances drawn from the test set, covering all six conversational conditions and both synthesis routes.

More information about the benchmark statistics can be found in Appendix A.5.

4 Experimental Setup

4.1 Evaluated Systems

We evaluate various transcript-based and direct-audio systems without ConvoAudit-specific fine-tuning, and further assess the effect of benchmark-specific training on three models:

- **Transcript-based systems.** Each transcript-based system pairs one ASR model with one LLM answerer, which answers questions based on the resulting transcript. In particular, we consider multiple ASR models, including Fun-ASR (An et al., 2025), Qwen3-ASR series (Shi et al., 2026), and Qwen-Audio-3.0-ASR series (Bian et al., 2026). The LLM answerers include DeepSeek-V4-Pro (Xu et al., 2026), GLM-5.2 (Zeng et al., 2026), Qwen3.7-Max (Qwen Team, 2026a), and Qwen3.8-Max (Qwen Team, 2026b). These combinations support comparisons of both transcription and answering components.
- **Direct-audio systems.** We evaluate both proprietary and open-weight models that answer questions directly from audio recordings, without relying on external transcripts. In particular, the proprietary systems include Gemini-3.1-Pro, Gemini-3.7-Flash and Qwen3.5-Omni series (Team, 2026), while the open-weight models include Baichuan-Omni-1.5 (Li et al., 2025), Kimi-Audio-7B-Instruct (Ding et al., 2025), MiMo-Audio-7B-Instruct (Zhang et al., 2025), MiniCPM-o-4.5 (Cui et al., 2026), MOSS-Audio-8B-Instruct (Yang et al., 2026), Qwen2.5-Omni and Qwen3-Omni series (Xu et al., 2025).
- **Benchmark-specific training.** We further fine-tune Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and Qwen3-Omni-30B-A3B-Instruct on the ConvoAudit training set. Through comparing each fine-tuned model with its corresponding base model on the test set, we evaluate the effects of benchmark-specific training.

More information about the configuration panel and inference settings can be found in Appendix B.1.

Table 1: Performance of representative transcript-based and direct-audio systems on ConvoAudit. We report the average QA scores for dialogue content (D), interaction structure (I), acoustic–temporal organization (A), and Overall on test and test-mini sets. For the transcript-based systems reported here, LLM answerers take transcripts generated by Qwen3-ASR-Flash as input. To facilitate comparison between the two pathways using the same answerer, Qwen3.5-Omni-Plus is evaluated separately with ASR transcripts and audio recordings. Complete results are provided in Appendix C.1.

		Test				**Test-mini**			
System	Interface	D	I	A	Overall	D	I	A	Overall
Transcript-based systems									
DeepSeek-V4-Pro	API	61.8	34.7	28.2	46.7	60.8	34.3	29.2	46.3
GLM-5.2	API	70.0	32.9	24.5	49.4	69.5	33.1	24.0	49.1
Qwen3.5-Omni-Plus (Transcript)	API	70.3	30.6	27.9	49.6	69.7	29.8	28.9	49.3
Qwen3.7-Max	API	68.6	31.9	34.4	50.5	67.5	31.7	34.5	49.9
Qwen3.8-Max	API	68.6	32.1	34.2	50.5	67.4	32.1	34.1	49.9
Direct-audio systems									
Kimi-Audio-7B-Instruct	Local	29.5	14.1	15.3	21.9	29.8	15.0	13.9	22.1
Baichuan-Omni-1.5	Local	28.7	16.3	19.8	23.2	29.3	18.0	18.4	23.6
MiniCPM-o-4.5	Local	40.6	16.8	25.1	30.3	40.9	18.3	24.6	30.8
MiMo-Audio-7B-Instruct	Local	42.5	22.7	22.3	32.4	42.2	23.8	21.7	32.4
MOSS-Audio-8B-Instruct	Local	43.0	20.6	35.6	34.8	43.3	21.0	36.0	35.1
Qwen3.5-Omni-Plus	API	69.7	46.3	58.2	60.3	69.4	45.4	58.5	60.0
Gemini-3.1-Pro	API	–	–	–	–	64.5	66.1	50.6	62.1
Gemini-3.7-Flash	API	–	–	–	–	72.6	87.7	68.6	76.2

4.2 Evaluation Protocol

As illustrated in Appendix A.5, ConvoAudit includes both choice-based and open-ended questions. Here, choice-based responses are scored using deterministic rules, while subjective responses are graded by Qwen3.7-Max. Besides, unparseable responses in deterministically scored tasks, as well as generation or grading failures, are assigned a score of zero.

In particular, we report the average QA score within each capability dimension and an Overall score across all evaluated QA instances. For system m, let $y_{mi} \in [0, 1]$ denote its normalized score on QA instance i, then the corresponding scores are defined as follows:

$$\mu_{md} = \frac{100}{|U_d|} \sum_{i \in U_d} y_{mi}, \qquad \text{Overall}_m = \frac{100}{|U|} \sum_{i \in U} y_{mi}, \tag{2}$$

where U_d is the set of evaluated QA instances in the capability dimension d, and U is the set of all evaluated QA instances. More information about evaluation protocol can be found in Appendix B.2.

5 Results and Discussion

We evaluate over 50 transcript-based and direct-audio system configurations, examining conversational conditions, synthesis-route and judge sensitivity, ASR effects on transcript-based systems, and benchmark-specific training, alongside human listening assessment. This section presents the main findings, with complete results and supplementary analyses provided in Appendix C.

5.1 Main Results

Table 1 summarizes the performance of representative transcript-based and direct-audio systems on test and test-mini across *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A). As presented in this table, individual systems exhibit substantial variation across

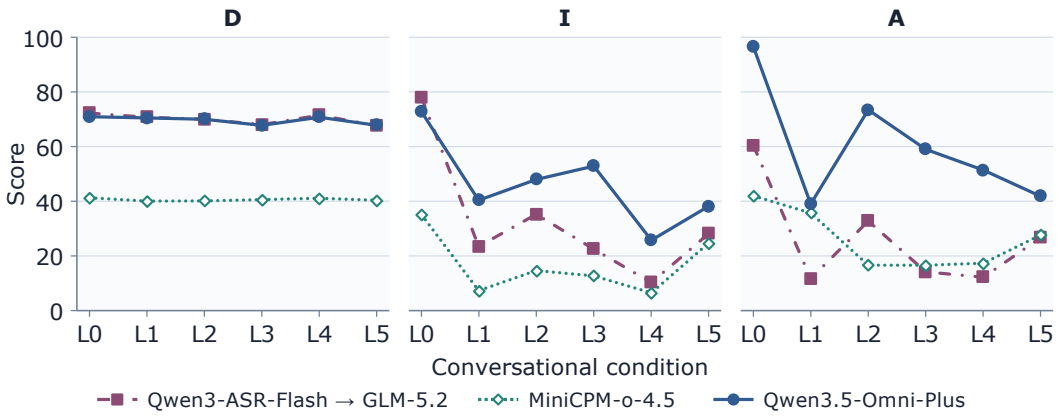


Figure 3: Performance of representative transcript-based and direct-audio systems across six conversational conditions (L0–L5). From left to right, the panels show average test QA scores for dialogue content (D), interaction structure (I), and acoustic–temporal organization (A). GLM-5.2 takes transcripts generated by Qwen3-ASR-Flash as input, while MiniCPM-o-4.5 and Qwen3.5-Omni-Plus process audio directly. Complete results are provided in Appendix C.2.

capability dimensions, and similar Overall scores can mask distinct capability profiles, highlighting the limitations of relying on Overall alone and motivate the need for dimension-specific evaluation.

This distinction is particularly obvious when Qwen3.5-Omni-Plus serves as the answerer in both pathways. On the test set, its transcript-based and direct-audio configurations achieve similar dialogue-content scores (70.3 versus 69.7). However, direct audio improves interaction structure from 30.6 to 46.3 and acoustic–temporal organization from 27.9 to 58.2. Comparable content recovery therefore does not imply comparable understanding of how a conversation unfolds.

The broader comparison further demonstrates the capability-specific strengths of direct-audio systems. For example, MOSS-Audio-8B-Instruct achieves 35.6 on acoustic–temporal organization, surpassing all listed transcript-based systems (24.5–34.4) despite their use of substantially larger LLM answerers. Qwen3.5-Omni-Plus also outperforms all listed transcript-based systems on both interaction structure and acoustic–temporal organization. These results suggest that direct audio access offers advantages beyond content recovery, enabling even a relatively compact model to capture acoustic–temporal information that larger transcript-based systems struggle to recover.

5.2 Capability profiles across conversational conditions

As described in Section 3.2, we construct six conversational conditions (L0–L5) that preserve the substantive mainline while varying background activity, side exchanges, and speaker transitions. Here, Figure 3 compares representative systems across these conditions to examine how conversational organization affects their capability profiles.

As presented in the figure, dialogue content remains relatively stable across conditions, whereas interaction structure and acoustic–temporal organization vary substantially. Despite similar dialogue-content scores, Qwen3.5-Omni-Plus outperforms transcript-based GLM-5.2 on both other dimensions under L1–L5. All three systems obtain their lowest interaction-structure scores under L4, which involves a temporary takeover followed by the original participant's return. Together, these results show that comparable content recovery can conceal substantial differences in understanding conversational organization. In particular, tracking participation and continuity through temporary speaker transitions remains challenging even when the mainline content is recoverable, underscoring the importance of evaluating capability profiles across conversational conditions.

5.3 Robustness across synthesis routes and judge models

To examine whether system comparisons depend on audio-synthesis and grading choices, we assess robustness across the two audio-synthesis routes and two judge models. As presented in Figure 4,

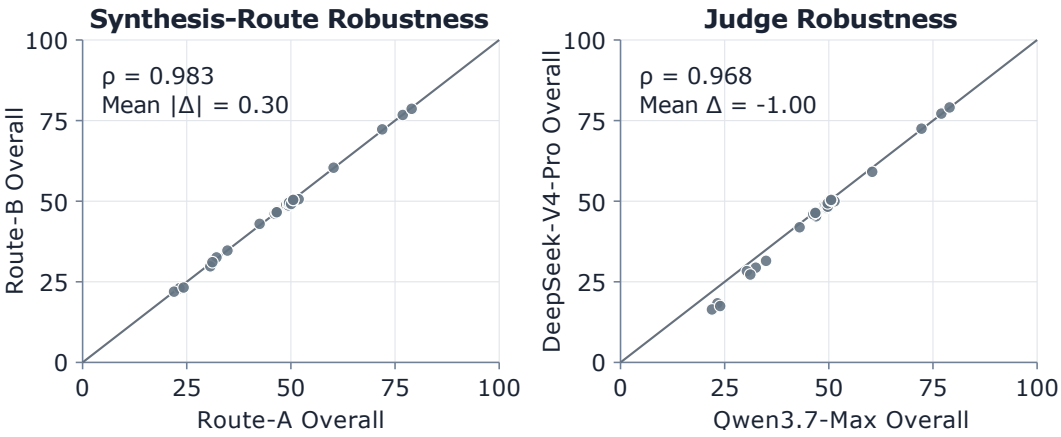


Figure 4: Consistency of test Overall scores across audio-synthesis routes (left) and judge models (right), where each point represents one system configuration. Diagonal lines indicate equal scores, and ρ denotes Spearman's rank correlation. Complete results are provided in Appendix C.3.

Table 2: Test Overall scores of LLM answerers using transcripts from different ASR models. Complete results are provided in Appendix C.4.

ASR Model	DeepSeek-V4-Pro	GLM-5.2	Qwen3.7-Max	Qwen3.8-Max
Fun-ASR	46.87	48.64	50.02	50.20
Qwen-Audio-3.0-ASR-Flash-Filetrans	46.54	48.82	50.09	50.04
Qwen-Audio-3.0-ASR-Flash	46.15	48.83	50.40	50.40
Qwen3-ASR-Flash-Filetrans	46.63	49.00	50.30	50.26
Qwen3-ASR-Flash	46.67	49.41	50.48	50.52

the two synthesis routes yield closely aligned scores and rankings, with a rank correlation of 0.983 and a mean absolute difference of only 0.30 points. Changing the judge also preserves broad ranking agreement ($\rho = 0.968$), although DeepSeek-V4-Pro assigns Overall scores that are 1.00 point lower on average. Together, these results suggest that our system comparisons remain largely consistent across the audio-synthesis routes and judge models.

5.4 Impact of ASR on transcript-based systems

We examine how ASR model selection and information supplied to LLM answerers affect transcript-based performance. As presented in Table 2, varying the ASR model while fixing the LLM answerer changes test Overall scores less than 1.00 points, indicating modest sensitivity to ASR model selection within the evaluated panel. We further examine structural evidence on the diagnostic subset while holding ASR words fixed. As presented in Table 3, adding timestamps to fixed ASR speech units substantially improves performance over words-only input, whereas adding speaker labels alone yields limited gains. However, when all fields are visible and reference-defined speech units are fixed, replacing either ASR timestamps or speaker labels with reference values improves performance. These findings suggest that the usefulness of structural evidence depends on both its availability and its correctness, highlighting the importance of preserving reliable temporal and speaker information alongside recognized words in transcript-based systems.

5.5 Effects of benchmark-specific training

To assess ConvoAudit's value for model improvement, we apply supervised fine-tuning (SFT) to three open-weight direct-audio models using its training set and compare their performance with that of their corresponding base models. As presented in Table 4, all three models improve substantially across dialogue content, interaction structure, and acoustic–temporal organization on the source-disjoint test set, with the largest gains in interaction structure. Notably, even the fine-tuned

Table 3: Effects of structural-evidence availability and correctness on transcript-based systems. We use Qwen-Audio-3.0-ASR-Flash-Filetrans outputs, entries report score changes on the evaluated diagnostic subset relative to the corresponding baselines. Complete results are provided in Appendix C.4.

Evidence Intervention	DeepSeek-V4-Pro	GLM-5.2	Qwen3.7-Max	Qwen3.8-Max
Availability: relative to words-only input				
Add timestamps	+18.66	+18.66	+17.22	+16.78
Add speaker labels	−0.27	+0.09	−1.08	+0.54
Correctness: relative to ASR-provided values				
Use reference timestamps	+4.32	+3.41	+4.44	+4.17
Use reference speaker labels	+5.23	+5.04	+8.25	+9.62

Table 4: Test QA scores before and after supervised fine-tuning (SFT) on ConvoAudit's training set. Complete results are provided in Appendix C.5.

	Before SFT				**After SFT**			
Model	D	I	A	Overall	D	I	A	Overall
Qwen2.5-Omni-3B	28.17	15.02	25.87	23.79	67.92	78.96	72.37	72.13
Qwen2.5-Omni-7B	35.31	23.23	32.47	31.13	71.77	85.77	76.17	76.85
Qwen3-Omni-30B-A3B-Instruct	49.65	31.29	43.23	42.85	76.02	82.42	80.62	78.89

Qwen2.5-Omni-3B outperforms the unadapted Qwen3.5-Omni-Plus baseline in Table 1, which is reported to have 397B parameters. The fine-tuned 3B model achieves the higher Overall score (72.13 versus 60.32), as well as higher scores on interaction structure and acoustic–temporal organization. These results highlight ConvoAudit's value for targeted model improvement, enabling compact models to outperform substantially larger proprietary baselines.

5.6 Human listening validation

We conduct the human listening assessment to examine audio clarity and the consistency of selected reference annotations, using recordings balanced across the two synthesis routes and six conversational conditions. In particular, audio is rated clear or mostly clear for 97.9% of the recordings, with human responses matching the reference answers on 91.5% of the answerable items. We further evaluate speaker annotations through anonymized comparisons between differing ASR and reference assignments with words and timestamps held fixed, finding that reference assignments are preferred in 95.8% of cases. These findings provide human evidence supporting the perceptual quality of the synthesized audio and the consistency of the reference annotations.

6 Conclusion

In this work, we introduced ConvoAudit, a controlled diagnostic benchmark for conversational audio understanding. It evaluates three complementary dimensions: dialogue content, interaction structure, and acoustic–temporal organization, revealing capability differences that aggregate scores can obscure. Shared-source variants preserve conversational mainlines across six conditions, including baseline dialogues, noninteractive background audio, side exchanges, persistent handoffs, temporary takeovers followed by returns, and their combinations, enabling controlled analysis of how systems' capability profiles vary with conversational organization. Our evaluation of over 50 system configurations demonstrates that similar dialogue-content performance can conceal substantial differences in interaction structure and acoustic–temporal organization, with the relative advantage of direct-audio systems varying across conditions. Matched transcript controls further show that the value of structural evidence depends on both its availability and its correctness. Fine-tuning also enables compact open-weight models to surpass strong proprietary baselines. Together, these findings demonstrate ConvoAudit's value for identifying condition-dependent capability gaps, examining which evidence improves system judgments, and guiding targeted model improvement.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript translation and language polishing. As described in the paper, the benchmark construction and evaluation pipelines use generative models for dialogue processing, question construction, audio synthesis, and grading of open-ended answers. AI models are also evaluated as candidate systems on the benchmark. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ConvoAudit is intended to support research, diagnostic evaluation, and targeted improvement of conversational audio understanding. Through controlled evaluation, we aim to help more researchers identify capability limitations and guide the development of more reliable systems. We do not endorse the use of the benchmark or associated resources for impersonation, non-consensual surveillance, or unsupported inferences about individuals. We encourage responsible use of these resources and interpretation of results within the scope of the evaluated tasks and conversational conditions.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe benchmark construction, quality checks, and data partitioning in Section 3, and the experimental setup in Section 4. Our experiments build on publicly available open-weight models and proprietary models accessed through public APIs, with model configurations and inference settings documented in Appendix B.1. Complete results and supplementary analyses are provided in Appendix C, and we will release the complete ConvoAudit benchmark upon acceptance to support independent evaluation and further research.

REFERENCES

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*, 2024.

Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, et al. Fun-asr technical report. *arXiv preprint arXiv:2509.12508*, 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Chuanmeng Bian, Daren Chen, Peixin Chen, Zhigao Chen, Zhiyun Fan, Zhifu Gao, Bo Gong, Qing Gu, Jiajun He, Yawei Hu, et al. Qwen-audio-3.0-asr technical report. *arXiv preprint arXiv:2609.07549*, 2026.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Yuhang Dai, Xin Shu, Zengxi Li, Lei Xie, Xiangang Li, and Jianwei Yu. Tag-bench: Benchmarking temporal audio grounding in large audio language models. *arXiv preprint arXiv:2609.01542*, 2026.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4763–4784, 2025.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, et al. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35740–35770, 2026.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. Natcs: Eliciting natural customer support dialogues. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9652–9677, 2023.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Kaiwen Luo, Zhenhong Zhou, Leyan Wang, Liang Lin, Tianyu Shao, Yuanhe Zhang, Yang Xiao, Yuxuan Li, Miao Yu, Kailin Lyu, et al. A survey of large audio language models: Generalization, trustworthiness, and outlook. *arXiv preprint arXiv:2605.20266*, 2026.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38, 2026.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457, 2021.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, et al. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Qwen Team. Qwen3.7: The agent frontier, May 2026a. URL `https://qwen.ai/blog?id=qwen3.7`.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026b. URL `https://qwen.ai/blog?id=qwen3.8`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Zhaokai Sun, Shuai Wang, Zhennan Lin, Chengyou Wang, Dehui Gao, Yuang Cao, Chunjiang He, Pan Zhou, and Lei Xie. Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios. *arXiv preprint arXiv:2606.22868*, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Thibaut Thonet, Laurent Besacier, and Jos Rozen. Elitr-bench: A meeting assistant benchmark for long-context language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 407–428, 2025.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026.

Bajian Xiang, Cheng Wen, Han Zhao, Hao Wang, Haoxu Wang, Jiawei Jin, Jiayan Cui, Jie Chen, Mengxi Nie, Tianyu Zhao, et al. Qwen-audio-3.0-tts: Freely controllable and highly robust speech synthesis with multi-stage training paradigm. *arXiv preprint arXiv:2607.23938*, 2026.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, et al. Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity. *arXiv preprint arXiv:2510.23541*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10155–10181, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, 2023.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, et al. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627, 2026.

A Benchmark Construction and Task Inventory

A.1 Source-dialogue curation

We curate multi-turn dialogue texts in Chinese and English from varied conversational settings, including customer service, peer discussions, and support interactions. Rule-based checks and model-assisted refinement standardize utterances, correct turn boundaries, and assess semantic coherence. Privacy review is also applied during curation. The resulting dialogues supply the substantive mainlines shared by the conversational variants.

To limit lexical overlap across splits, we normalize the dialogue text, form character 5-shingles, and identify near duplicates using a Jaccard-similarity threshold of 0.5. All variants derived from a dialogue, and all dialogues in a detected near-duplicate group, remain in the same split. This lexical check reduces cross-split overlap but cannot rule out semantic similarity or exposure during model training.

Recording counts, audio duration, and QA coverage are reported in Appendix A.5.

A.2 Conversational variant construction

Table 5 defines the six shared-source conditions. A persistent handoff and a temporary takeover/return can involve identical voice counts but differ in whether the original voice resumes.

Table 5: Six conversational conditions preserving the shared source mainline. L1–L4 are separate variants, and L5 combines selected components.

Condition	Conversational organization
L0: Mainline	Main dialogue with native paralinguistic events.
L1: Background	Noninteractive background audio accompanies the mainline.
L2: Aside	A local side exchange occurs, followed by mainline resumption.
L3: Handoff	A new voice takes over one main channel and continues.
L4: Takeover/return	A new voice temporarily takes over; the original voice returns.
L5: Composition	Two to four components from L1–L4 occur together.

Mainline preservation concerns substantive dialogue content. It does not hold all auxiliary utterances, background layers, prosody, or timing fixed. L0 may already contain backend-native nonlexical vocal events; L1–L4 independently add the specified organization, and L5 combines selected components. Events inherited from L0 are distinguished from the added condition-specific events.

A.3 Dual-route audio synthesis

Route A jointly generates long-form audio with SoulX-Podcast-1.7B (Xie et al., 2025); Route B synthesizes individual turns with Qwen-Audio-3.0-TTS-Flash (Xiang et al., 2026) and composes them into the complete recording.

Route A applies voice-group batching when required for joint long-form synthesis. Background mixing, auxiliary exchanges, and event composition realize the conditions; unchanged turns can reuse L0 audio. The records distinguish logical conversational roles from physical voices, allowing a new voice to occupy an existing role. The synthesis implementations use shared role/voice assignments and event specifications, but retain separate audio, timelines, and event records.

Conversational mainlines and event specifications remain matched across two audio-synthesis routes. Their separately realized records provide the reference evidence for each recording and support the cross-implementation consistency checks in Appendix C.3.1.

A.4 Question–answer pair construction and quality checks

Dialogue-content questions and reference answers are generated with constrained prompts grounded in the cleaned conversational mainline and reused across its variants. Interaction-structure and

acoustic–temporal questions and references are derived deterministically from the voices, scenes, events, and timelines realized in each recording. Task-specific checks verify response-format compliance, supporting evidence, reference consistency under option reordering, and audio–timeline alignment. Construction metadata are excluded from standard model inputs.

These automatic checks establish consistency between questions, references, and their supporting records. Realized timelines describe speech units and registered events; they do not exhaustively transcribe background layers. An utterance window containing a vocal event therefore need not identify the event's exact acoustic onset and offset.

For eligible instances of task t, question/reference construction $\mathcal{G}_t$ uses source-mainline evidence u_s for dialogue content (D); interaction structure (I) and acoustic–temporal organization (A) additionally use realized record z_{scr}:

$$(q_{scrt}, a^{\star}_{scrt}) = \mathcal{G}_t(E_{scrt}), \qquad E_{scrt} = \begin{cases} u_s, & t \in \mathrm{D}, \\ (u_s, z_{scr}), & t \in \mathrm{I} \cup \mathrm{A}. \end{cases} \tag{3}$$

Table 6 numbers tasks within each reporting dimension: D-1–D-7 for dialogue content, I-1–I-4 for interaction structure, and A-1–A-3 for acoustic–temporal organization. The examples in Appendix A.5 illustrate the question and response format of each task.

Table 6: Test-set task inventory for dialogue content (D), interaction structure (I), and acoustic–temporal organization (A). Structured D responses use rubric-based judging; choice responses and structured I-2/A-3 responses use deterministic scoring.

Dim.	Task	Reference target	Response	QA items
D	D-1	Explicit facts, entities, constraints, and states	Structured	1,152
D	D-2	Functions and relations of anchored utterances	Multi-choice	984
D	D-3	Intent ownership, alternatives, final requests	Structured	1,176
D	D-4	Text-supported attitude trajectory	Multi-choice	1,200
D	D-5	Roles, actions, authority, responsibility	Structured	1,200
D	D-6	Dialogue phases, progress, unresolved actions	Structured	1,152
D	D-7	Rule-constrained conclusions from distributed evidence	Structured	1,104
I	I-1	Turn-forming voices and their composition	Single-choice	1,200
I	I-2	Additional source, channel, mainline continuity	Structured	1,200
I	I-3	Earliest extra interaction and entry phase	Single-choice	1,200
I	I-4	Extra speech and its relation to the mainline	Single-choice	1,200
A	A-1	Voice attributes and nonlexical vocal events	Single-choice	1,200
A	A-2	Global audible structure of the exchange	Single-choice	1,200
A	A-3	Condition-specific event timing and windows	Structured	1,017

Eligibility is determined from the required source and event evidence before candidate scoring. A source lacking the evidence required by a semantic template does not receive that question; a temporal question similarly requires the appropriate realized reference. Thus, 1,200 test recordings do not imply 1,200 eligible items for every task. The inventory contains 5,784 open structured D items, 2,184 D choice items, and 8,217 I/A items with deterministic contracts. The dimensions contain 7,968 D, 4,800 I, and 3,417 A items. Missing references preclude instance construction. For eligible instances, unparseable responses in deterministic tasks, generation failures, and grading failures receive zero and remain in the applicable denominator.

Dialogue-content formats. D-1 requests facts within two named facets, with utterance anchors and reconciliation of updates. D-2 selects the supported dialogue-act judgments for two anchored utterances and their relation. D-3 reports an intent trajectory or the ownership/status of candidate requests. D-4 selects text-supported statements about attitude progression; prosody is outside its reference contract. D-5 reports roles, duties, and authority boundaries or applies explicit service rules. D-6 reports resolved/conditional/unresolved issues and next actions, or ordered functional phases with textual anchors. D-7 provides evidence-supported reasoning and a conclusion chosen from `holds`, `does_not_hold`, and `cannot_be_determined`. The structured variants retain their task-specific schemas and reference rubrics.

Interaction formats. I-1 inventories distinct voices that form conversational turns within a specified window; unaddressed background chatter does not count as an interacting voice. I-2 has four fields: presence of an additional speaking source, primary channel, mainline continuity, and the perceived gender of the earliest additional interacting voice. Channels distinguish background speech, side exchange, persistent handoff, and temporary takeover; continuity distinguishes unchanged, resumed, and persistently changed mainlines. I-3 selects the earliest extra interacting voice's perceived gender and entry in the first third versus the later two-thirds, with separate choices for noninteractive extra speech or its absence. I-4 identifies how additional speech relates semantically and interactively to the mainline. These voice descriptors concern the synthesis/reference contract and do not infer a person's gender identity.

Acoustic–temporal formats. A-1 jointly identifies the reference voice attribute and nonlexical vocal-event state of anchored speech. A-2 selects the overall audible structure, including a complete component combination for L5. A-3 uses the temporal granularity supported by the realized records (Table 7); planned timestamps are not reference boundaries. Of its 1,017 test items, 817 request numeric windows and 200 L1 items request coarse temporal patterns. The numeric subset comprises 17 L0 items and 200 items in each of L2–L5. A-1 and A-2 each contribute 200 L0 items. Consequently, the A dimension contains 417 L0 items and 600 items in each other condition; condition comparisons also reflect this difference in reference composition.

Table 7: Temporal-reference formats for acoustic–temporal organization (A), task A-3. Utterance windows do not assert exact acoustic event boundaries.

Condition	Test items	Required reference granularity
L0	17	Utterance windows containing registered paralinguistic events.
L1	200	Whole-recording background patterns: continuous, intermittent, or sparse.
L2	200	Aside entry, response, and mainline-resumption utterance windows.
L3	200	Handoff and new-voice entry windows and their ordering.
L4	200	Temporary entry window, takeover interval, original-voice return window.
L5	200	Component-specific events with waveform-occurrence or utterance-window boundaries.

A.5 Benchmark statistics and examples

Table 8 reports the benchmark's recording counts, audio duration, and QA counts. The 7,929 recordings comprise 3,974 Route-A and 3,955 Route-B realizations. Validation and test include all six conditions and both synthesis routes for every included dialogue. Training includes incomplete panels where a realization is unavailable; missing recordings are not replaced by silence. All variants of a dialogue remain in the same split across both synthesis routes.

Table 8: Benchmark statistics by split. Audio duration is summed over realized recordings, and QA counts include separate questions attached to the same recording.

Quantity	Training	Validation	Test	Total
Recordings	5,769	960	1,200	7,929
Duration (hours)	348.88	61.45	74.30	484.63
QA instances	76,956	12,787	16,185	105,928

The test set contains 74.30 hours of audio and 16,185 QA instances: 7,968 for dialogue content (49.2%), 4,800 for interaction structure (29.7%), and 3,417 for acoustic–temporal organization (21.1%). Chinese and English questions concern distinct conversations rather than translations of the same dialogue. A question's language does not imply that every utterance in its recording is monolingual.

For rapid preliminary evaluation, test-mini provides a subset of 3,237 test QA instances covering all six conversational conditions and both synthesis routes. Appendix B.2 details its evaluation scope.

Examples across the fourteen tasks. The following held-out examples provide one question for every task, including its answer choices or structured-response requirements and its expected answer. Tasks concerning the same recording appear together. Gray transcript excerpts provide reading context and are not additional model input. Descriptive speaker aliases replace internal labels consistently in the excerpts and displayed answers. Choice sets, answer facts, and event times are preserved from the evaluation examples.

Shared instruction for the direct-audio examples: "You are an audio question-answering assistant. Answer only from the audio provided by the user; be accurate and concise, and strictly follow the format required by the question."

Movie discussion with a local aside

Paper-only transcript excerpt *not model input; [. . .] marks omissions*

Voice 1 [0.000–2.658 s]: What is the title of this movie?
Voice 2 [2.658–4.758 s]: It's called Real Steel.
[. . .] **Voice 2** [14.178–18.758 s]: The movie is based on a short story called Steel, written by Richard Matheson.
[. . .] **Voice 2** [21.898–27.348 s]: Yes, it's about the story called Steel, published in May nineteen fifty-eight.
[. . .] **Voice 1** [32.908–36.822 s]: Who are the actors in it, and where is the story based?
Voice 2 [36.822–41.722 s]: Hugh Jackman stars as Charlie, and the story is based in the U.S.
Additional voice [41.722–45.822 s]: Excuse me, do you know where the nearest printer is?
Voice 1 [45.822–48.962 s]: It's down the hall, second door on the left.
Voice 1 [48.962–50.962 s]: Sorry about that. Let's continue.
[. . .] **Voice 1** [59.522–62.356 s]: What time period is the movie set in?
Voice 2 [62.356–65.336 s]: It's set in twenty twenty. [. . .]

D-1 / Explicit facts Structured response

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. Exhaustively extract every explicit discourse fact in the "work title" and "release year" dimensions.

Response format. Return one JSON object with exactly five keys: entities, constraints, reconciliation, final_facts, and item_count. entities, constraints, and final_facts are arrays of objects containing exactly item:string and anchor:string (anchor: at most 8 words); reconciliation is a non-empty string (use a complete-scope absence statement when no reconciliation is established); item_count is an integer equal to the combined number of items in entities, constraints, and final_facts. Do not add keys.

EXPECTED OUTPUT / D-1

```
{
  "entities": [
    {"item": "Real Steel", "anchor": "It's called Real Steel"},
    {"item": "Steel", "anchor": "story called Steel"}
  ],
  "constraints": [
    {"item": "May nineteen fifty-eight", "anchor": "published in
     May nineteen fifty-eight"}
  ],
  "reconciliation": "The dialogue explicitly identifies 'Real
   Steel' as the movie title and 'Steel' as the source short
```

```
    story; the release year 'May nineteen fifty-eight' is stated
    specifically for the short story 'Steel', not the movie, and
    no conflicting dates or titles appear elsewhere in the
    conversation.",
  "final_facts": [],
  "item_count": 3
}
```


I-1 / Turn-taking voice inventory — Select one

MODEL INPUT **Question**

Across the whole call, how many distinct turn-taking voices are there, and what is their gender composition? Count only voices that speak turns which get responded to or that respond to others; background chatter that no one answers does not count. Select the single best option.

Options:

A. exactly 3 turn-taking voices: 1 male-sounding and 2 female-sounding

B. exactly 2 turn-taking voices: 1 male-sounding and 1 female-sounding

C. exactly 3 turn-taking voices: 2 male-sounding and 1 female-sounding

D. exactly 4 turn-taking voices: 2 male-sounding and 2 female-sounding

E. exactly 3 turn-taking voices: 0 male-sounding and 3 female-sounding

F. exactly 2 turn-taking voices: 0 male-sounding and 2 female-sounding

EXPECTED OUTPUT / I-1

```
A
```

I-2 / Additional speech and mainline continuity — Structured response

MODEL INPUT **Question**

First locate the exchange whose core discussion includes "story called Steel, published in May nineteen fifty-eight". Determine whether all intelligible speech remains with the opening sources or an additional source speaks. Summarize whether a speaking source is added, its primary channel type, and whether the mainline resumes. Intelligible spoken content is required; a scene layer or non-lexical cue does not qualify.

Response format. Return exactly a JSON object with four keys: additional_speaking_source_present, primary_channel, mainline_continuity, and additional_source_gender. additional_speaking_source_present is boolean; primary_channel must be one of opening_mainline_only, background_speech, side_exchange, persistent_handoff, temporary_takeover; mainline_continuity must be one of unchanged, resumed, persistently_changed; additional_source_gender is the perceived gender of the earliest additional turn-taking speaking source and must be one of male, female, none (use none when no additional turn-taking source exists; background chatter that never forms a turn counts as none). Do not add keys.

EXPECTED OUTPUT / I-2


```
{
  "additional_speaking_source_present": true,
  "primary_channel": "side_exchange",
  "mainline_continuity": "resumed",
```

```
  "additional_source_gender":  "female"
}
```


I-4 / Extra speech and the mainline — Select one

MODEL INPUT **Question**

Review any speech content beyond the two mainline parties in this recording. Which option is correct? Judge by the semantic relation between the extra speech content and the mainline business, and by whether it interacts with the mainline. Select the single best option.

Options:

A. Apart from the two mainline parties' speech and their occasional non-lexical vocal sounds, no additional speech content exists

B. Background voices exist whose content is close to the mainline topic or everyday chatter, but they never interact with the mainline

C. An additional voice speaks content directly relevant to the mainline business and persistently carries one party's task to the end

D. An additional voice speaks content directly relevant to the mainline business as a temporary supplement, then hands back to the original speaker

E. A brief extra exchange occurs whose content is unrelated to the mainline business, after which the mainline resumes

F. Background voices exist and directly participate in the mainline business discussion

EXPECTED OUTPUT / I-4

```
E
```

A-3 / Event time windows — Structured response

Reading note (not model input): Here, "after" means after the selected mainline utterance; the side exchange is followed by mainline resumption.

MODEL INPUT **Question**

In the call whose main discussion includes "Who are the actors in it, and where is the story based?", locate every brief foreground side exchange: its audible entry, response, and mainline resumption. Give utterance anchors and realized start/end seconds in chronological order.

Response format. Return one JSON object and no unlisted keys. Fields and types: side_exchanges is a non-empty array and granularity:string. Each exchange contains exactly insertion_timing_relative_to_mainline (one of before, mid, after), side_entry, side_response, and mainline_resumption. Each of the last three uses the anchor shape {speaker:string,utterance:string,utterance_window_seconds:{start_sec:number,end_sec:number}}.

EXPECTED OUTPUT / A-3


```
{
  "side_exchanges":  [
    {
      "insertion_timing_relative_to_mainline":  "after",
      "side_entry":  {
        "speaker":  "Additional voice",
```

```
        "utterance": "Excuse me, do you know where the nearest
          printer is?",
        "utterance_window_seconds": {"start_sec": 41.72,
          "end_sec": 45.82}
      },
      "side_response": {
        "speaker": "Voice 1",
        "utterance": "It's down the hall, second door on the
          left.",
        "utterance_window_seconds": {"start_sec": 45.82,
          "end_sec": 48.96}
      },
      "mainline_resumption": {
        "speaker": "Voice 1",
        "utterance": "Sorry about that. Let's continue.",
        "utterance_window_seconds": {"start_sec": 48.96,
          "end_sec": 50.96}
      }
    }
  ],
  "granularity": "realized utterance windows"
}
```


Account verification and unresolved follow-up

Paper-only transcript excerpt *not model input; [. . .] marks omissions*

[. . .] **Customer** [3.212–4.972 s]: I want to verify my account.
[. . .] **Agent** [7.652–11.454 s]: May I know what function you want to use for Alipay certification?
Customer [11.954–13.554 s]: Passport with face scan.
[. . .] **Agent** [25.568–28.928 s]: May I have the last four digits of your certified documents?
[. . .] **Customer** [33.928–35.928 s]: One two three four.
[. . .] **Agent** [40.788–44.868 s]: I have checked that you have already attempted to scan your face, right?
Customer [45.368–47.755 s]: Yes, but my account verification failed.
Customer [48.255–50.291 s]: I don't know what the reason behind this was.
[. . .] **Agent** [52.571–55.876 s]: I will now submit an application to help investigate further.
Customer [56.376–58.835 s]: OK, I am waiting for your response.
Agent [59.335–63.151 s]: I am conducting further inquiries for you. Please be patient.

D-2 / Utterance functions and relations Select all that apply

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. One line in the dialogue says: "I will now submit an application to help investigate further." A later line says: "I am conducting further inquiries for you. Please be patient." Select every supported statement about the pragmatic function of these utterances and their relation.

Options:

A. The two utterances function as mutually exclusive alternatives where the second retracts the proposal made in the first.

B. The second utterance describes an ongoing procedural activity rather than confirming a final resolution.

C. The two utterances jointly establish a workflow continuation where the second reports progress on the commitment made in the first.

D. The first utterance confirms that the account verification failure has been fully resolved.

E. The first utterance commits to initiating a review process without guaranteeing a specific outcome.

F. The second utterance serves as a definitive notice that the investigation is complete.

EXPECTED OUTPUT / D-2

```
B, C, E
```

A-1 / Voice attributes and vocal events Select one

MODEL INPUT **Question**

A quality check needs one clip verified: for the voice that says "I will now submit an application to help investigate further.", which description matches the audio? Select the single best option.

Options:

A. a female-sounding voice accompanied by yawning

B. a male-sounding voice accompanied by yawning

C. a female-sounding voice accompanied by sighing

D. a male-sounding voice accompanied by sighing

E. a male-sounding voice with no non-lexical vocal sound

F. a female-sounding voice with no non-lexical vocal sound

EXPECTED OUTPUT / A-1

```
F
```

D-3 / Intent trajectory Structured response

Reading note (not model input): The focal speaker is the customer seeking account verification.

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. Based on the dialogue, analyze the focal speaker's intent trajectory and provide a structured response.

Response format. Return one JSON object with exactly five fields and no unlisted fields. Required JSON value types: "surface_question":non-empty string; "underlying_goal":non-empty string; "abandoned_or_retained_alternative":non-empty string; "final_requested_action":non-empty string; "convergence_evidence":non-empty array of non-empty strings. Field roles: surface_question is the focal speaker's explicit question; underlying_goal is the broader goal; abandoned_or_retained_alternative records alternatives and their disposition; final_requested_action is the last explicitly requested or accepted action; convergence_evidence supports whether the goal converged or remains unresolved. If any of the first four facts is not explicitly established, put a clear

natural-language absence statement in that string field; do not use null, an object, an array, or an empty string.

EXPECTED OUTPUT / D-3

```
{
  "surface_question": "No explicit interrogative question was
    asked by the focal speaker in the terminal exchange; the
    speaker instead made a declarative statement of ignorance
    regarding the failure reason.",
  "underlying_goal": "The focal speaker aims to successfully
    complete the account verification process using passport and
    face scan, which was initially stated as 'I want to verify my
    account' and reiterated through cooperation with security
    checks.",
  "abandoned_or_retained_alternative": "No alternative
    verification method or workaround is explicitly proposed,
    rejected, or retained by the focal speaker throughout the
    dialogue; the interaction remains focused solely on resolving
    the failure of the originally selected passport with face scan
    method.",
  "final_requested_action": "The focal speaker explicitly accepts
    the agent's proposal to submit an investigation application by
    responding 'OK, I am waiting for your response', thereby
    requesting the agent to proceed with that specific
    investigative action and provide subsequent feedback.",
  "convergence_evidence": [
    "Customer states: 'Yes, but my account verification
      failed.'",
    "Customer states: 'I don't know what the reason behind this
      was.'",
    "Agent states: 'I will now submit an application to help
      investigate further.'",
    "Customer responds: 'OK, I am waiting for your response.'"
  ]
}
```

D-4 / Text-supported stance

Select all that apply

MODEL INPUT Question

Answer from the principal conversational thread identified by the question. Using only the wording across the dialogue, which of the following statements avoid confusing the early stance, later adjustment, and final explicit position? Use at least three separated statements and test the strongest competing interpretation. Differences between the candidate descriptions must be settled by the dialogue wording itself. Select every fully supported statement from the structured choices.

Options:

A. The speaker concludes with an active inquiry stance by stating they are conducting further inquiries and asking the customer to be patient.

B. The speaker initially adopts a resolution stance by promising to fix the account verification failure immediately after receiving document details.

C. Upon learning that the customer's face scan failed for an unknown reason, the speaker shifts to a directive stance by instructing the customer to retry the face scan immediately.

D. The speaker initially adopts a procedural verification stance by requesting specific document details to enable further checking.

E. The speaker concludes with a completed resolution stance by confirming that the investigation application has been successfully processed and the issue is resolved.

F. Upon learning that the customer's face scan failed for an unknown reason, the speaker shifts to an investigative stance by submitting an application to investigate further.

EXPECTED OUTPUT / D-4

```
A, D, F
```

D-6 / Progress and pending actions — Structured response

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. Track questions and owned commitments, requests, accepted assignments, conditions, and closure requirements; ignore assertions, opinions, or trivia. resolved: directly answered questions. unresolved: unanswered or uncertain-estimate questions. conditional and completion_guard need links. next_actions: pending owned commitments, requests, or assignments; remove only when the exact scope is explicitly completed, not after a related update. Exclude unowned suggestions. Return only resolved, conditional, unresolved, next_actions, completion_guard. Values are {"status":"established","items":[non-empty strings]} or, when empty, {"status":"none_explicitly_established","items":[]}.

EXPECTED OUTPUT / D-6


```
{
  "resolved": {
    "status": "established",
    "items": [
      "What function to use for Alipay certification (answered:
       Passport with face scan)",
      "Last four digits of certified documents (answered: One two
       three four)",
      "Whether user has already attempted to scan face (answered:
       Yes)"
    ]
  },
  "conditional": {"status": "none_explicitly_established",
    "items": []},
  "unresolved": {"status": "established", "items": ["Reason
    behind account verification failure"]},
  "next_actions": {"status": "established", "items": ["Submit
    application to investigate further", "Conduct further
    inquiries"]},
  "completion_guard": {"status": "established", "items": ["User
    waiting for response regarding investigation outcome"]}
}
```

Movie discussion: roles and evidence

Paper-only transcript excerpt *not model input; [. . .] marks omissions*

[. . .] **Voice 1** [2.100–5.460 s]: The movie is Batman versus Superman. Have you seen it?
Voice 2 [5.960–10.600 s]: No, I haven't. I am a fan of Batman though. What is it about?
[. . .] **Voice 1** [13.600–20.960 s]: It's about Superman coming back to life and having to face Batman while they are both in a state of confusion.
[. . .] **Voice 2** [30.400–34.160 s]: How does Superman come back to life? And a state of confusion?
[. . .] **Voice 1** [41.820–45.066 s]: I don't want to say much more so as not to spoil it.
[. . .] **Voice 1** [48.019–53.145 s]: It's one hundred and fifty-three minutes long, and there is some good comedy in it.
[. . .] **Voice 1** [63.405–67.245 s]: I'll say this about it, they bring kryptonite back in the mix.
Voice 1 [67.745–72.046 s]: I'm not sure if I like the actor that plays Superman. [. . .]

D-5 / Roles and responsibility Structured response

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. Based on the dialogue, analyze the functional relationship between the two participants and provide a structured assessment covering identities, actual duties, authority boundaries, escalation or non-decidable items, and responsibility attribution. Workflow actions like explaining plot details do not constitute formal duties or authority unless explicitly assigned.

Response format. Return one JSON object with exactly five keys: identities, actual_duties, authority_boundaries, escalation_or_non_decidable_items, responsibility_attribution. Each value is an object with exactly status and items. status is established when items is a non-empty array of non-empty strings; status is not_formally_established exactly when items is an empty array. identities must be established with a non-empty factual description of both participants. Do not add keys.

EXPECTED OUTPUT / D-5

```
{
  "identities": {
    "status": "established",
    "items": [
      "Voice 1 acts as a movie recommender providing detailed plot
       and cast information about Batman versus Superman",
      "Voice 2 acts as an information seeker asking clarifying
       questions about the movie's content and genre"
    ]
  },
  "actual_duties": {"status": "not_formally_established",
   "items": []},
  "authority_boundaries": {"status": "not_formally_established",
   "items": []},
  "escalation_or_non_decidable_items": {"status":
   "not_formally_established", "items": []},
  "responsibility_attribution": {"status":
   "not_formally_established", "items": []}
}
```

D-7 / Evidence and conclusion Structured response

MODEL INPUT **Question**

Answer from the principal conversational thread identified by the question. Complete claim under test: The movie is explicitly confirmed to contain good comedy and kryptonite, and the speaker expresses unreserved positive sentiment toward it.

Decision rules. Let P be the complete claim and decide it only from the complete principal conversational thread, scanned from beginning to end. Preserve every predicate of P, including subject, scope, current applicability, modality, outcome, and quantifier, and reconcile later updates before deciding. holds means that the complete thread supplies sufficient positive evidence for every predicate of P and every required applicability/bridge rule, with no sufficient counterevidence left unresolved after temporal reconciliation. does_not_hold means that the complete thread supplies sufficient negative or contradictory evidence for not-P, with no sufficient positive evidence left unresolved after temporal reconciliation. cannot_be_determined means that both sides are insufficient, or that sufficient positive and negative evidence remain in unresolved conflict at call end. The mere absence of a positive promise, confirmation, or outcome is never negative evidence and must not yield does_not_hold. Absence may count negatively only when P itself explicitly defines a closed, exhaustively enumerated scope in which that absence is the negation.

Response format. Return one JSON object with exactly hops, conclusion, conclusion_basis, and counterfactual_check. hops is an ordered list of at least 3 necessary, non-duplicate objects, each containing exactly evidence_quote:string and inference:string. Each evidence_quote is copied verbatim from one utterance and uses at most 8 English words. Every hop must use a different utterance, and the hops follow those utterances' order; each inference must advance the decision about P, not merely repeat or decorate another hop. conclusion is exactly holds, does_not_hold, or cannot_be_determined; conclusion_basis is one non-empty sentence applying this truth table; counterfactual_check is one non-empty string changing exactly one explicit premise. Do not add keys.

EXPECTED OUTPUT / D-7


```
{
  "hops": [
    {"evidence_quote": "there is some good comedy", "inference":
      "The speaker explicitly confirms the presence of good comedy
      in the movie."},
    {"evidence_quote": "bring kryptonite back", "inference":
      "The speaker explicitly confirms that kryptonite is included
      in the movie."},
    {"evidence_quote": "not sure if I like", "inference": "The
      speaker expresses uncertainty about the lead actor,
      contradicting unreserved positive sentiment."}
  ],
  "conclusion": "does_not_hold",
  "conclusion_basis": "Although comedy and kryptonite are
    confirmed, the speaker's expressed uncertainty about the actor
    refutes the predicate of unreserved positive sentiment.",
  "counterfactual_check": "If the speaker had expressed certainty
    about liking the actor, the claim of unreserved positive
    sentiment would hold given the other confirmed predicates."
}
```

Account verification with a persistent handoff

Paper-only transcript excerpt *not model input; [. . .] marks omissions*

Agent [0.000–3.212 s]: Hello, this is Alex. How can I help you?
Customer [3.212–5.472 s]: I want to verify my account.
[. . .] **Agent** [40.788–48.888 s]: I understand you'd like this done quickly, let me transfer you to our supervisor who can look into this further for you, please hold on.
New agent [50.888–57.868 s]: Hello, I'm the supervisor here, my colleague has already briefed me on your account verification situation.
[. . .] **New agent** [80.228–83.908 s]: I am conducting further inquiries for you. Please be patient.

I-3 / Additional voice entry Select one

MODEL INPUT **Question**

Use "I am conducting further inquiries for you. Please be patient" to identify the call. For the earliest voice outside the opening participants that forms an interactive turn (if any), what is its perceived gender and does it first appear in the first third or later two thirds of the call? Select the single best option.

Options:

A. A male-sounding voice outside the opening participants forms its earliest interactive turn in the first third of the call

B. A male-sounding voice outside the opening participants forms its earliest interactive turn in the later two thirds of the call

C. A female-sounding voice outside the opening participants forms its earliest interactive turn in the first third of the call

D. A female-sounding voice outside the opening participants forms its earliest interactive turn in the later two thirds of the call

E. A voice outside the opening participants is audible but never forms an interactive turn (no one responds)

F. No voice outside the opening participants is audible in the call

EXPECTED OUTPUT / I-3

B

Account verification in a composed scene

Paper-only transcript excerpt *not model input; [. . .] marks omissions*

Agent [0.000–3.212 s]: Hello, this is Alex. How can I help you?
Customer [3.212–5.472 s]: I want to verify my account.
[. . .] **Agent** [40.788–43.688 s]: Let me grab a colleague on this, hold on a sec.
New agent [45.688–51.788 s]: Hi, I'm taking over from here. My colleague filled me in on your situation, so let's continue.
[. . .] **New agent** [71.988–75.188 s]: I am conducting further inquiries for you. Please be patient.

Paper-only scene note. The realized record also registers a noninteractive office-equipment background layer. The transcript excerpt does not encode that acoustic evidence.

A-2 / Complete audible scene Select one

MODEL INPUT **Question**

Which option gives the complete set of sound structures simultaneously realized in this composite recording? Select the single best option.

Options:

A. Present: a non-interactive background sound layer; a new voice persistently taking over one side to the end; a temporary voice takeover followed by the original voice returning. Absent: a brief side exchange followed by mainline resumption.

B. Present: a new voice persistently taking over one side to the end. Absent: a non-interactive background sound layer; a brief side exchange followed by mainline resumption; a temporary voice takeover followed by the original voice returning.

C. Present: a non-interactive background sound layer; a brief side exchange followed by mainline resumption; a new voice persistently taking over one side to the end. Absent: a temporary voice takeover followed by the original voice returning.

D. Present: a non-interactive background sound layer; a new voice persistently taking over one side to the end. Absent: a brief side exchange followed by mainline resumption; a temporary voice takeover followed by the original voice returning.

E. Present: a non-interactive background sound layer. Absent: a brief side exchange followed by mainline resumption; a new voice persistently taking over one side to the end; a temporary voice takeover followed by the original voice returning.

F. Present: none beyond the opening mainline exchange. Absent: a non-interactive background sound layer; a brief side exchange followed by mainline resumption; a new voice persistently taking over one side to the end; a temporary voice takeover followed by the original voice returning.

EXPECTED OUTPUT / A-2

D

B Supplementary Experimental Setup

This appendix supplements Section 4, and Appendix C provides supplementary results.

B.1 Evaluated systems and configurations

Configuration inventory. The evaluation covers 52 configurations: 50 on test and two Gemini endpoints on test-mini only. Detailed results cover 43 test systems and the two Gemini endpoints. Table 10 lists 40 unadapted test systems and both Gemini endpoints; Appendix C.5 compares three SFT models with their bases. Systems are grouped by input pathway, transcript-based or direct-audio; API/Local denotes access mode. Shared lineages, answerers, and ASR frontends make configurations dependent. Additional routes, tasks, judges, and diagnostic interventions do not increase this inventory.

The reported direct-audio panel spans Baichuan, Gemini, Kimi, MiMo, MiniCPM, MOSS, and Qwen. The reported transcript-based panel combines five ASR interfaces with six answerers. The interfaces are Fun-ASR, Qwen-Audio-3.0-ASR-Flash, Qwen-Audio-3.0-ASR-Flash-Filetrans, Qwen3-ASR-Flash, and Qwen3-ASR-Flash-Filetrans. The answerers are DeepSeek-V4-Pro, GLM-5.2, Qwen3.5-Omni-Plus, Qwen3.5-Plus, Qwen3.7-Max, and Qwen3.8-Max. These configurations do not constitute a representative sample of available models.

Recorded inference settings. All 43 reported test configurations use temperature 0 and a 2,048-token output budget on both synthesis routes.

The Qwen3.8-Max diagnostics use these settings with thinking disabled. They reuse recorded ASR and reference traces and report deterministic task scores. CORE analyses evaluate 2,217 test QA

instances spanning scene recognition (A-2) and temporal localization (A-3); task-specific controls retain their corresponding subsets. The visibility, field-quality, timestamp, speaker, lexical-source, and representation controls are detailed in Appendix C.4. These interventions do not add leaderboard configurations.

Recorded training setup. Table 9 gives the training recipes for the three SFT models compared with their bases. Training and validation use their respective splits across both synthesis routes. The audio tower, vision tower, and speech-output modules remain frozen. Shared settings are global batch size 48, micro batch size 1, Adam, bfloat16, weight decay 0.1, cosine learning-rate decay, 5% warmup, and seed 42. Minimum learning rates are 10^{-6} for the two Qwen2.5 runs and 10^{-7} for Qwen3. Scheduled epochs are training budgets; evaluation uses intermediate training states, with no documented common selection criterion.

Table 9: Training recipes for the paired SFT comparisons. Maximum length is the sequence limit.

Model	Learning Rate	Max Length	Scheduled Epochs
Qwen2.5-Omni-3B SFT	10^{-5}	9,344	1
Qwen2.5-Omni-7B SFT	10^{-5}	9,344	1
Qwen3-Omni-30B-A3B-Instruct SFT	3×10^{-6}	15,360	1

B.2 Evaluation protocol supplements

Evaluation scope and failures. Test is the primary evaluation scope: 16,185 QA instances attached to 1,200 recordings. Supplementary analyses use test or explicitly defined task subsets. Every eligible instance is retained, with generation and grading failures scored at zero (Appendix C.3.2). Missing references instead preclude instance construction.

Nested test-mini evaluation. Test-mini contains 3,237 test QA instances attached to 1,147 recordings: 1,618 and 1,619 instances from Routes A and B, covering all six conditions. Membership follows the original test-mini evaluations; the sampling algorithm and seed are not inferred. Each instance matches test in recording, reference, choices, base question, answer type, response contract, and language. Test-mini scores for test-evaluated systems restrict their existing outputs to this subset.

Gemini-3.7-Flash and Gemini-3.1-Pro each have 3,237 predictions from separate test-mini runs; their 1,161 subjective-answer instances use Qwen3.7-Max. Only 1,720 complete question strings match the test prompts verbatim despite matching base questions. Gemini results thus differ in inference scope and prompt wording from restricted test outputs. These test-mini-only direct-audio systems have missing test entries and are not ranked against test scores.

Metrics and deterministic contracts. Equation 2 averages QA scores within *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A), and across all evaluated QA instances for Overall. Every eligible instance receives equal weight within its reported scope. Consequently, Overall is the QA-count-weighted mean of the three dimension scores, not their unweighted mean. For test results, these implicit weights follow the QA proportions in Figure 2; test-mini and diagnostic subsets use their own eligible QA proportions. The following contracts define the underlying item scores.

For choice tasks, the scorer applies exact matching to parsed option-index sets. Both D-2 (984 test items) and D-4 (1,200 test items) are multi-selection tasks: a subset, extra option, or unparseable answer receives zero. Other choice tasks follow their recorded answer type. I-2 requires valid JSON with exactly four fields: additional-source presence, primary channel, mainline continuity, and additional-source gender. All four values must match; field-wise accuracy is a diagnostic, not the official I-2 score.

A-3 applies condition-specific contracts. For ordinary windows, let K and J be required and predicted interval counts, e_b the error of required boundary b, and m_b indicate a valid matched boundary. Its

score is

$$y_{\text{win}} = \frac{1}{2K} \sum_{b=1}^{2K} m_b \max\left(0, 1 - \frac{|e_b|}{10\,\text{s}}\right) \min\left(1, \frac{K}{\max(1, J)}\right). \tag{4}$$

Missing boundaries contribute zero; extra intervals incur the count penalty. L5 additionally gates each event's boundaries on its canonical component, event type, and boundary granularity at the prescribed position; exact anchor wording is diagnostic only. L1 instead exactly matches the unordered multiset of coarse temporal patterns. Consequently, A-3 comprises 817 numeric-window items and 200 L1 pattern items, including only 17 eligible L0 items. It is not a uniform timestamp-error metric. Subjective D-1/D-3/D-5/D-6/D-7 answers, including structured responses, use rubric judgments on a 0–4 overall scale, normalized by four; the overall judgment is not a mechanical mean of rubric dimensions. Audited hard-constraint violations cap the normalized subjective score at 0.5 rather than automatically forcing zero.

Primary-judge setup. All 43 reported test configurations use Qwen3.7-Max for subjective-answer grading. Five transcript-based configurations, one per ASR frontend, also use Qwen3.7-Max as their answerer. This shared identity can affect subjective comparisons, while deterministic scores are unaffected. Appendix C.3.2 evaluates sensitivity to an alternative judge with candidate answers held fixed. A common judge standardizes scoring but does not establish judge–answerer independence.

Source-cluster resampling. Point estimates average eligible QA instances. For uncertainty estimates, we resample source dialogues as clusters, retaining their conditions, synthesis routes, and QA instances together. Each bootstrap draw recomputes the score as the resampled score sum divided by the resampled QA count, scaled to 0–100; paired comparisons share the same draws. The 95% percentile intervals use 10,000 draws and quantify source-sampling variation conditional on the recorded predictions and references, not training-seed or endpoint-version variability. Seeds are 1024 for condition profiles and matched-pathway contrasts, 512 for route comparisons, and 256 for evidence diagnostics and adaptation analyses.

C Complete Experimental Results

This appendix expands the experimental results in Section 5, including the complete main-system comparison, conversational conditions, robustness analyses, structural-evidence interventions, and training results. The system inventory, training setup, scoring contracts, and resampling procedures are specified in Appendix B.

C.1 Complete main results

The expanded inventory includes every system and score cell shown in Table 1, reporting *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A), alongside Overall. Transcript-based systems are grouped by their actual ASR frontend. Configurations with test results are ordered by ascending Overall within each frontend group and within the direct-audio category; test-mini-only rows follow without implying a test rank.

Table 10: Complete main-system results on test and test-mini, supplementing Table 1. D, I, A, and Overall are QA-average scores; API/Local denotes the answering interface. Each system has separate test and test-mini rows.

System	Interface	Scope	D	I	A	Overall
Transcript-based systems (Fun-ASR)						
DeepSeek-V4-Pro	API	test	62.5	34.4	28.1	46.9
		test-mini	62.1	33.9	29.0	46.7
GLM-5.2	API	test	68.9	32.6	23.9	48.6
		test-mini	68.9	33.8	23.4	48.9

Continued on next page

System	Interface	Scope	D	I	A	Overall
Transcript-based systems (Fun-ASR) (continued)						
Qwen3.5-Plus	API	test	67.0	31.6	31.7	49.0
		test-mini	66.7	30.8	31.3	48.6
Qwen3.5-Omni-Plus	API	test	69.6	30.6	27.9	49.2
		test-mini	68.8	30.1	28.5	48.8
Qwen3.7-Max	API	test	67.8	31.9	34.1	50.0
		test-mini	66.3	32.0	34.6	49.4
Qwen3.8-Max	API	test	68.1	31.8	34.3	50.2
		test-mini	66.8	31.5	34.4	49.5
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash)						
DeepSeek-V4-Pro	API	test	61.5	34.1	27.3	46.2
		test-mini	60.0	34.0	27.6	45.4
GLM-5.2	API	test	69.3	32.6	23.8	48.8
		test-mini	69.5	32.7	24.6	49.1
Qwen3.5-Plus	API	test	67.2	31.2	30.9	48.9
		test-mini	66.8	31.0	30.6	48.6
Qwen3.5-Omni-Plus	API	test	69.7	30.5	28.0	49.3
		test-mini	69.2	30.3	29.2	49.2
Qwen3.7-Max	API	test	68.2	32.3	34.2	50.4
		test-mini	67.2	31.9	35.0	49.9
Qwen3.8-Max	API	test	68.4	32.4	33.8	50.4
		test-mini	66.9	32.0	34.1	49.6
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash-Filetrans)						
DeepSeek-V4-Pro	API	test	61.7	34.5	28.1	46.5
		test-mini	61.0	33.2	28.3	45.8
Qwen3.5-Plus	API	test	67.0	31.1	31.3	48.8
		test-mini	66.5	30.7	29.7	48.1
GLM-5.2	API	test	69.0	33.3	23.5	48.8
		test-mini	68.0	33.9	22.9	48.4
Qwen3.5-Omni-Plus	API	test	69.5	30.5	27.6	49.1
		test-mini	69.0	30.1	29.1	49.0
Qwen3.8-Max	API	test	67.9	31.9	33.9	50.0
		test-mini	66.8	31.7	34.8	49.6
Qwen3.7-Max	API	test	67.9	32.2	33.7	50.1
		test-mini	66.9	31.7	33.8	49.5
Transcript-based systems (Qwen3-ASR-Flash)						
DeepSeek-V4-Pro	API	test	61.8	34.7	28.2	46.7
		test-mini	60.8	34.3	29.2	46.3
GLM-5.2	API	test	70.0	32.9	24.5	49.4
		test-mini	69.5	33.1	24.0	49.1
Qwen3.5-Omni-Plus	API	test	70.3	30.6	27.9	49.6
		test-mini	69.7	29.8	28.9	49.3
Qwen3.5-Plus	API	test	68.4	31.0	32.1	49.6
		test-mini	68.1	30.1	31.5	49.1
Qwen3.7-Max	API	test	68.6	31.9	34.4	50.5
		test-mini	67.5	31.7	34.5	49.9
Qwen3.8-Max	API	test	68.6	32.1	34.2	50.5
		test-mini	67.4	32.1	34.1	49.9

Continued on next page

System	Interface	Scope	D	I	A	Overall
Transcript-based systems (Qwen3-ASR-Flash-Filetrans)						
DeepSeek-V4-Pro	API	test	61.7	34.7	28.2	46.6
		test-mini	61.4	35.1	29.0	46.7
GLM-5.2	API	test	69.6	32.6	23.8	49.0
		test-mini	69.4	32.8	23.9	48.9
Qwen3.5-Plus	API	test	67.2	31.2	32.0	49.1
		test-mini	67.4	30.5	31.8	48.9
Qwen3.5-Omni-Plus	API	test	70.1	30.6	27.5	49.4
		test-mini	69.5	29.9	27.4	48.9
Qwen3.8-Max	API	test	68.3	31.9	34.0	50.3
		test-mini	67.5	31.0	34.0	49.6
Qwen3.7-Max	API	test	68.1	32.2	34.1	50.3
		test-mini	67.1	32.6	34.2	49.9
Direct-audio systems						
Kimi-Audio-7B-Instruct	Local	test	29.5	14.1	15.3	21.9
		test-mini	29.8	15.0	13.9	22.1
Baichuan-Omni-1.5	Local	test	28.7	16.3	19.8	23.2
		test-mini	29.3	18.0	18.4	23.6
Qwen2.5-Omni-3B	Local	test	28.2	15.0	25.9	23.8
		test-mini	28.1	15.4	24.9	23.6
MiniCPM-o-4.5	Local	test	40.6	16.8	25.1	30.3
		test-mini	40.9	18.3	24.6	30.8
Qwen2.5-Omni-7B	Local	test	35.3	23.2	32.5	31.1
		test-mini	35.4	24.8	30.7	31.3
MiMo-Audio-7B-Instruct	Local	test	42.5	22.7	22.3	32.4
		test-mini	42.2	23.8	21.7	32.4
Direct-audio systems (continued)						
MOSS-Audio-8B-Instruct	Local	test	43.0	20.6	35.6	34.8
		test-mini	43.3	21.0	36.0	35.1
Qwen3-Omni-30B-A3B-Instruct	Local	test	49.7	31.3	43.2	42.9
		test-mini	49.9	31.5	43.0	43.0
Qwen3.5-Omni-Flash	API	test	55.7	34.2	44.1	46.9
		test-mini	55.1	33.4	42.7	46.1
Qwen3.5-Omni-Plus	API	test	69.7	46.3	58.2	60.3
		test-mini	69.4	45.4	58.5	60.0
Gemini-3.1-Pro	API	test	–	–	–	–
		test-mini	64.5	66.1	50.6	62.1
Gemini-3.7-Flash	API	test	–	–	–	–
		test-mini	72.6	87.7	68.6	76.2

Scores for all fourteen tasks. Tables 11 and 12 report all fourteen task scores for 45 configurations, covering *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A). Each task score is the mean over its evaluated QA instances. Eligibility varies by task and condition. In particular, pooled A-3 combines numeric and coarse-pattern contracts and must not be read as uniform temporal-boundary accuracy.

The 43 test configurations include the three SFT models reported in Appendix C.5 and use the same frozen outputs and judge as the dimension-level comparison. The two Gemini configurations are evaluated on test-mini only; the Scope column makes this distinction explicit. Qwen3.7-Max serves as both answerer and subjective-answer judge for its transcript-based configurations, a dependence examined in Appendix C.3.2.

Table 11: Dialogue-content task scores (D-1–D-7). Each score averages evaluated QA instances. Scope distinguishes test from test-mini-only configurations; API/Local denotes the answering interface.

System	Interface	Scope	D-1	D-2	D-3	D-4	D-5	D-6	D-7
Transcript-based systems (Fun-ASR)									
DeepSeek-V4-Pro	API	test	49.5	78.3	68.1	64.2	77.7	55.1	45.3
GLM-5.2	API	test	56.9	84.5	74.2	89.5	76.0	56.5	45.1
Qwen3.5-Plus	API	test	54.5	87.8	65.9	86.2	68.8	58.3	48.6
Qwen3.5-Omni-Plus	API	test	54.9	89.8	68.8	91.9	77.7	55.9	48.8
Qwen3.7-Max	API	test	56.6	89.9	75.5	76.2	72.5	57.5	47.8
Qwen3.8-Max	API	test	58.4	90.5	75.0	76.1	72.9	58.0	47.7
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash)									
DeepSeek-V4-Pro	API	test	48.7	75.5	68.2	60.4	76.8	54.6	46.7
GLM-5.2	API	test	56.9	84.6	74.8	88.5	76.0	58.0	46.5
Qwen3.5-Plus	API	test	55.1	87.6	65.8	86.8	68.0	58.2	50.6
Qwen3.5-Omni-Plus	API	test	55.9	88.6	69.2	92.0	77.2	56.0	50.1
Qwen3.7-Max	API	test	57.9	90.4	75.5	76.3	72.7	57.3	49.0
Qwen3.8-Max	API	test	57.9	90.4	75.4	76.4	72.8	57.9	49.8
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash-Filetrans)									
DeepSeek-V4-Pro	API	test	49.6	77.2	66.9	60.2	77.7	55.7	45.4
Qwen3.5-Plus	API	test	54.5	87.6	65.7	87.8	68.0	58.3	48.6
GLM-5.2	API	test	56.2	86.4	74.0	87.8	75.6	58.1	45.5
Qwen3.5-Omni-Plus	API	test	55.4	89.7	68.7	92.5	77.1	55.2	48.4
Qwen3.8-Max	API	test	57.4	90.0	74.7	75.8	72.7	57.9	48.6
Qwen3.7-Max	API	test	58.2	89.6	75.0	75.9	72.8	57.7	47.8
Transcript-based systems (Qwen3-ASR-Flash)									
DeepSeek-V4-Pro	API	test	50.6	76.0	66.9	59.8	77.0	56.6	46.6
GLM-5.2	API	test	57.5	85.8	73.3	90.9	76.7	59.9	46.3
Qwen3.5-Omni-Plus	API	test	56.1	89.2	68.5	92.4	76.8	59.4	50.6
Qwen3.5-Plus	API	test	56.5	87.5	66.2	89.5	69.1	61.0	49.9
Qwen3.7-Max	API	test	58.7	89.8	75.1	76.5	72.7	58.9	49.9
Qwen3.8-Max	API	test	59.2	90.3	74.4	76.6	72.4	59.7	49.4
Transcript-based systems (Qwen3-ASR-Flash-Filetrans)									
DeepSeek-V4-Pro	API	test	49.0	77.0	67.3	60.1	77.7	55.6	46.3
GLM-5.2	API	test	56.6	84.9	74.2	90.4	76.4	59.0	46.0
Qwen3.5-Plus	API	test	54.0	87.0	66.1	87.4	68.7	58.8	49.5
Qwen3.5-Omni-Plus	API	test	55.5	89.3	68.5	93.0	77.4	57.1	50.4
Qwen3.8-Max	API	test	57.8	89.5	75.6	77.7	72.6	57.9	48.4
Qwen3.7-Max	API	test	57.6	89.1	75.6	76.2	73.0	57.8	49.0
Direct-audio systems									
Kimi-Audio-7B-Instruct	Local	test	35.2	19.7	35.8	4.1	47.3	31.6	31.6
Baichuan-Omni-1.5	Local	test	34.4	23.9	33.8	3.7	38.5	35.6	31.3
Qwen2.5-Omni-3B	Local	test	28.5	26.4	32.0	19.0	47.1	19.6	23.7
MiniCPM-o-4.5	Local	test	36.1	50.5	43.7	23.4	49.4	43.1	39.4
Qwen2.5-Omni-7B	Local	test	27.5	39.5	36.0	24.9	55.0	30.2	34.2
MiMo-Audio-7B-Instruct	Local	test	43.9	25.3	46.1	33.0	67.9	39.1	38.7

Continued on next page

System	Interface	Scope	D-1	D-2	D-3	D-4	D-5	D-6	D-7
Direct-audio systems (continued)									
MOSS-Audio-8B-Instruct	Local	test	35.5	54.5	39.7	41.5	61.8	32.7	36.0
Qwen3-Omni-30B-A3B-Instruct	Local	test	47.9	42.2	51.3	46.7	65.8	49.3	42.4
Qwen3.5-Omni-Flash	API	test	48.8	59.2	52.4	64.0	68.3	52.8	43.2
Qwen3.5-Omni-Plus	API	test	53.6	90.4	68.3	92.3	76.6	57.3	50.0
Qwen2.5-Omni-3B-SFT	Local	test	56.2	90.1	57.1	86.2	82.4	58.5	46.1
Qwen2.5-Omni-7B-SFT	Local	test	62.1	94.6	60.3	92.4	82.9	61.7	49.7
Direct-audio systems (continued)									
Qwen3-Omni-30B-A3B-Instruct-SFT	Local	test	67.3	96.0	68.1	97.3	83.6	64.2	56.7
Gemini-3.1-Pro	API	test-mini	37.2	83.3	66.1	85.8	66.0	49.9	66.0
Gemini-3.7-Flash	API	test-mini	46.8	91.1	73.3	88.3	70.2	58.4	83.6

Table 12: Interaction-structure (I-1–I-4) and acoustic–temporal task scores (A-1–A-3). Each score averages evaluated QA instances. Scope distinguishes test from test-mini-only configurations; API/Local denotes the answering interface.

System	Interface	Scope	I-1	I-2	I-3	I-4	A-1	A-2	A-3
Transcript-based systems (Fun-ASR)									
DeepSeek-V4-Pro	API	test	45.2	20.8	26.8	44.6	41.5	37.2	1.4
GLM-5.2	API	test	41.4	22.2	15.2	51.4	35.1	32.4	0.7
Qwen3.5-Plus	API	test	45.8	17.8	17.3	45.3	47.1	37.5	6.8
Qwen3.5-Omni-Plus	API	test	46.0	15.3	14.1	47.2	41.8	36.6	1.2
Qwen3.7-Max	API	test	45.9	15.3	16.8	49.6	54.8	39.3	3.4
Qwen3.8-Max	API	test	45.7	15.3	16.2	49.8	55.2	39.3	3.6
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash)									
DeepSeek-V4-Pro	API	test	44.8	20.8	26.2	44.8	40.2	36.6	1.3
GLM-5.2	API	test	41.8	21.5	15.5	51.8	34.2	32.8	1.1
Qwen3.5-Plus	API	test	44.2	18.1	16.8	45.9	45.5	36.9	6.7
Qwen3.5-Omni-Plus	API	test	47.0	14.6	13.8	46.7	42.9	36.1	0.8
Qwen3.7-Max	API	test	45.9	16.0	17.8	49.6	54.8	39.8	3.5
Qwen3.8-Max	API	test	46.0	16.2	17.3	49.8	55.0	38.7	3.1
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash-Filetrans)									
DeepSeek-V4-Pro	API	test	45.2	20.6	28.2	43.8	42.1	36.8	1.4
Qwen3.5-Plus	API	test	44.5	18.0	16.5	45.3	46.7	36.4	7.0
GLM-5.2	API	test	42.4	24.6	14.6	51.6	33.9	32.4	0.8
Qwen3.5-Omni-Plus	API	test	46.0	15.0	14.2	46.8	41.9	35.8	1.1
Qwen3.8-Max	API	test	45.9	14.9	17.0	49.6	54.7	39.0	3.5
Qwen3.7-Max	API	test	46.3	15.5	16.8	50.0	54.1	38.9	3.5
Transcript-based systems (Qwen3-ASR-Flash)									
DeepSeek-V4-Pro	API	test	45.8	20.8	27.9	44.2	41.9	37.2	1.4
GLM-5.2	API	test	42.2	22.2	15.2	52.0	35.2	33.8	0.8
Qwen3.5-Omni-Plus	API	test	46.5	14.7	14.7	46.7	43.0	35.7	1.0
Qwen3.5-Plus	API	test	44.3	17.5	16.8	45.5	48.2	37.3	6.8
Qwen3.7-Max	API	test	45.7	15.8	17.2	49.0	55.7	39.2	3.5
Qwen3.8-Max	API	test	45.7	15.9	17.2	49.7	55.8	38.5	3.5

Continued on next page

System	Interface	Scope	I-1	I-2	I-3	I-4	A-1	A-2	A-3
Transcript-based systems (Qwen3-ASR-Flash-Filetrans)									
DeepSeek-V4-Pro	API	test	46.3	20.9	27.7	43.9	42.4	36.7	1.3
GLM-5.2	API	test	41.8	22.2	15.2	51.2	34.8	32.4	0.7
Qwen3.5-Plus	API	test	44.3	18.1	17.0	45.5	48.3	37.2	6.5
Qwen3.5-Omni-Plus	API	test	46.5	14.9	13.9	46.9	41.5	35.9	1.1
Qwen3.8-Max	API	test	46.0	15.4	16.8	49.5	55.0	39.0	3.2
Qwen3.7-Max	API	test	46.2	15.8	16.8	50.2	55.0	39.4	3.2
Direct-audio systems									
Kimi-Audio-7B-Instruct	Local	test	14.2	3.7	14.9	23.5	24.9	18.0	0.9
Baichuan-Omni-1.5	Local	test	22.5	3.5	21.0	18.2	35.3	19.9	1.2
Qwen2.5-Omni-3B	Local	test	22.3	1.3	12.7	23.8	47.8	25.8	0.0
MiniCPM-o-4.5	Local	test	13.2	9.6	10.7	33.8	46.4	24.4	0.9
Qwen2.5-Omni-7B	Local	test	23.4	8.7	19.6	41.2	59.5	32.3	0.7
MiMo-Audio-7B-Instruct	Local	test	30.6	8.6	14.6	37.1	33.1	29.5	1.0
Direct-audio systems (continued)									
MOSS-Audio-8B-Instruct	Local	test	18.4	10.1	12.4	41.3	75.6	24.8	1.2
Qwen3-Omni-30B-A3B-Instruct	Local	test	39.9	24.3	18.0	42.9	85.8	31.0	7.5
Qwen3.5-Omni-Flash	API	test	50.9	20.7	21.2	44.2	83.1	33.3	10.8
Qwen3.5-Omni-Plus	API	test	60.8	27.8	29.7	66.9	91.9	41.2	38.5
Qwen2.5-Omni-3B-SFT	Local	test	69.4	73.8	79.0	93.6	87.9	92.8	29.9
Qwen2.5-Omni-7B-SFT	Local	test	78.4	80.4	88.4	95.8	93.8	95.4	32.6
Direct-audio systems (continued)									
Qwen3-Omni-30B-A3B-Instruct-SFT	Local	test	76.8	76.5	81.9	94.5	98.5	90.5	47.9
Gemini-3.1-Pro	API	test-mini	73.8	58.3	47.5	85.0	73.8	57.1	15.7
Gemini-3.7-Flash	API	test-mini	91.7	80.8	82.5	95.8	88.3	80.8	31.0

C.2 Conversational conditions

Figure 3 presents point estimates for *dialogue content* (D), *interaction structure* (I), and *acoustic–temporal organization* (A). For the matched Qwen3.5-Omni-Plus endpoint pooled over conditions, the direct-audio-minus-transcript differences are -0.67 points for D (95% CI $[-1.46, 0.10]$), $+15.69$ for I $[14.02, 17.40]$, and $+30.32$ for A $[27.61, 33.05]$.

Table 13 reports all three dimension estimates and their pointwise source-cluster intervals for every condition shown in Figure 3. It includes the figure's Qwen3.5-Omni-Plus and MiniCPM-o-4.5 direct-audio systems and Qwen3-ASR-Flash → GLM-5.2 transcript baseline, together with supplementary MiMo-Audio-7B-Instruct results.

Table 13: Complete condition profiles for the three systems in Figure 3, plus MiMo-Audio-7B-Instruct. D, I, and A are QA-average test scores; brackets give pointwise 95% source-cluster intervals, pooling both routes.

Condition	D [95% CI]	I [95% CI]	A [95% CI]
	Qwen3.5-Omni-Plus		
L0	70.93 [69.09, 72.73]	72.75 [69.25, 76.25]	96.52 [94.86, 97.97]
L1	70.43 [68.73, 72.04]	40.38 [38.00, 42.88]	39.00 [36.17, 42.00]
L2	70.11 [68.35, 71.84]	48.00 [44.62, 51.50]	73.38 [69.79, 76.94]
L3	67.77 [65.98, 69.57]	52.88 [49.12, 56.62]	58.99 [56.71, 61.19]
L4	70.82 [68.98, 72.58]	25.75 [21.12, 30.62]	51.29 [47.80, 54.85]
L5	67.88 [66.04, 69.67]	38.12 [34.62, 41.75]	41.89 [38.82, 45.04]

Continued on next page

Condition	D [95% CI]	I [95% CI]	A [95% CI]
		MiMo-Audio-7B-Instruct	
L0	43.17 [41.07, 45.35]	45.00 [41.00, 48.88]	50.12 [45.08, 55.18]
L1	42.60 [40.30, 45.06]	14.75 [12.75, 16.75]	27.33 [23.83, 30.83]
L2	41.89 [39.72, 44.12]	28.00 [25.25, 30.75]	17.77 [14.41, 21.20]
L3	42.90 [40.87, 44.97]	18.62 [16.25, 21.12]	13.37 [10.46, 16.28]
L4	42.32 [40.34, 44.39]	9.75 [7.75, 11.75]	16.06 [12.89, 19.22]
L5	42.04 [39.77, 44.37]	20.12 [17.38, 22.88]	17.53 [14.37, 20.69]
		MiniCPM-o-4.5	
L0	41.28 [38.93, 43.59]	35.12 [32.38, 38.00]	41.97 [35.90, 48.10]
L1	40.06 [38.02, 42.07]	7.25 [5.50, 9.25]	35.83 [32.33, 39.33]
L2	40.19 [37.97, 42.45]	14.62 [12.38, 16.88]	16.67 [13.79, 19.57]
L3	40.53 [38.19, 42.88]	12.75 [10.25, 15.25]	16.59 [13.56, 19.65]
L4	41.08 [38.75, 43.45]	6.50 [4.88, 8.25]	17.24 [14.48, 19.96]
L5	40.29 [38.22, 42.32]	24.62 [22.50, 26.88]	27.67 [24.34, 31.00]
		Qwen3-ASR-Flash → GLM-5.2	
L0	72.25 [70.55, 73.87]	78.00 [75.25, 80.75]	60.43 [55.98, 64.88]
L1	70.84 [69.06, 72.55]	23.38 [21.62, 25.25]	11.67 [9.33, 14.00]
L2	69.94 [68.00, 71.82]	35.12 [32.88, 37.50]	32.83 [29.08, 36.52]
L3	68.03 [66.33, 69.69]	22.50 [18.75, 26.12]	14.12 [11.36, 16.99]
L4	71.54 [69.61, 73.43]	10.25 [7.50, 13.25]	12.28 [9.74, 14.90]
L5	67.68 [65.60, 69.67]	28.12 [25.75, 30.50]	26.68 [23.14, 30.22]

Table 14 pairs this endpoint across direct audio and lexical Qwen3-ASR-Flash input within each condition. Each condition contains 1,328 D and 800 I QA instances. A has 417 at L0 and 600 elsewhere. This eligibility mixture means pooled A need not equal an unweighted mean of its six condition scores. The 18 condition-specific intervals are pointwise, with no simultaneous-coverage claim.

Table 14: Matched-endpoint condition contrasts: direct Qwen3.5-Omni-Plus minus Qwen3-ASR-Flash → Qwen3.5-Omni-Plus. Differences are in QA-average score points; brackets give paired, pointwise 95% source-cluster intervals.

Condition	Dimension	Direct	ASR	Difference	95% CI
L0	D	70.93	70.93	+0.00	[-1.26, 1.26]
L0	I	72.75	76.75	-4.00	[-7.00, -1.12]
L0	A	96.52	67.15	+29.37	[24.75, 33.93]
L1	D	70.43	71.48	-1.05	[-2.33, 0.19]
L1	I	40.38	23.62	+16.75	[14.38, 19.12]
L1	A	39.00	13.17	+25.83	[22.00, 29.67]
L2	D	70.11	70.59	-0.49	[-1.87, 0.93]
L2	I	48.00	29.62	+18.38	[14.75, 22.00]
L2	A	73.38	42.92	+30.46	[26.25, 34.69]
L3	D	67.77	68.19	-0.41	[-1.73, 0.89]
L3	I	52.88	22.88	+30.00	[25.50, 34.50]
L3	A	58.99	15.00	+43.99	[40.35, 47.66]
L4	D	70.82	71.44	-0.62	[-1.82, 0.57]
L4	I	25.75	8.12	+17.62	[13.38, 22.25]
L4	A	51.29	15.26	+36.03	[31.77, 40.44]
L5	D	67.88	69.35	-1.47	[-2.58, -0.34]
L5	I	38.12	22.75	+15.38	[11.75, 19.00]
L5	A	41.89	25.96	+15.93	[11.81, 20.15]

C.3 SYNTHESIS-ROUTE AND JUDGE SENSITIVITY

C.3.1 SYNTHESIS-ROUTE CONSISTENCY

We reaggregate the recorded scores for the 43 test configurations using the same Qwen3.7-Max subjective-answer judge as Table 1, reporting *dialogue content* (D), *interaction structure* (I), *acoustic–temporal organization* (A), and Overall. Route A contributes 8,093 QA instances and Route B 8,092. Their intersection contains 8,089 matched QA pairs; the seven unmatched instances all concern L0 temporal localization. Route membership is identical across configurations. The matched analysis removes this eligibility discrepancy while preserving the source, task, question case, and condition pairing. Each score averages its eligible QA instances. The audit uses 10,000 source-cluster bootstrap draws with seed 1024; identical draws are applied jointly to both routes and answering pathways. Every draw recomputes scores from the resampled score sums and QA counts.

Table 15 gives the configuration-level scores underlying the route comparison in Figure 4. It reports both routes for every test configuration, rather than only the configuration-level summary statistics.

Table 15: Complete test scores by synthesis route for all 43 reported configurations, including the three SFT models. D, I, A, and Overall are QA-item averages on each route's recorded eligibility set, using the primary Qwen3.7-Max subjective-answer judge. Systems follow the frontend groups and pooled test ordering of the main comparison. Each system has separate Route A and Route B rows.

System	Interface	Route	D	I	A	Overall
Transcript-based systems (Fun-ASR)						
DeepSeek-V4-Pro	API	A	62.09	34.88	27.68	46.75
		B	62.86	33.83	28.44	46.99
GLM-5.2	API	A	68.97	33.00	24.35	48.88
		B	68.91	32.12	23.46	48.41
Qwen3.5-Plus	API	A	66.73	31.62	33.26	49.25
		B	67.18	31.50	30.23	48.80
Qwen3.5-Omni-Plus	API	A	69.36	31.21	28.29	49.37
		B	69.77	30.08	27.48	49.07
Qwen3.7-Max	API	A	67.84	31.42	34.62	50.02
		B	67.71	32.42	33.50	50.02
Qwen3.8-Max	API	A	67.86	31.08	34.99	50.01
		B	68.41	32.42	33.60	50.39
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash)						
DeepSeek-V4-Pro	API	A	61.53	34.42	26.64	46.12
		B	61.42	33.83	28.03	46.19
GLM-5.2	API	A	69.42	33.04	23.82	49.00
		B	69.22	32.21	23.85	48.67
Qwen3.5-Plus	API	A	67.12	31.29	31.86	49.05
		B	67.30	31.17	30.00	48.71
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash) (continued)						
Qwen3.5-Omni-Plus	API	A	69.85	30.83	28.28	49.50
		B	69.63	30.21	27.67	49.08
Qwen3.7-Max	API	A	68.33	31.62	34.62	50.33
		B	68.08	33.04	33.88	50.47
Qwen3.8-Max	API	A	68.61	31.87	33.91	50.39
		B	68.18	32.83	33.71	50.42

Continued on next page

System	Interface	Route	D	I	A	Overall
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash-Filetrans)						
DeepSeek-V4-Pro	API	A	61.33	35.33	26.97	46.37
		B	62.07	33.58	29.32	46.71
Qwen3.5-Plus	API	A	66.82	31.21	32.59	49.03
		B	67.23	30.96	29.91	48.60
GLM-5.2	API	A	68.80	33.17	24.20	48.81
		B	69.25	33.42	22.89	48.83
Qwen3.5-Omni-Plus	API	A	69.52	30.54	27.86	49.16
		B	69.41	30.42	27.42	48.98
Qwen3.8-Max	API	A	67.63	31.37	34.22	49.83
		B	68.19	32.33	33.65	50.26
Qwen3.7-Max	API	A	67.80	31.67	33.92	49.93
		B	68.05	32.67	33.47	50.26
Transcript-based systems (Qwen3-ASR-Flash)						
DeepSeek-V4-Pro	API	A	61.88	35.12	27.00	46.58
		B	61.76	34.25	29.36	46.76
GLM-5.2	API	A	69.94	33.04	24.85	49.48
		B	70.15	32.75	24.17	49.35
Transcript-based systems (Qwen3-ASR-Flash) (continued)						
Qwen3.5-Omni-Plus	API	A	70.20	30.54	28.10	49.55
		B	70.46	30.71	27.73	49.65
Qwen3.5-Plus	API	A	68.64	31.17	32.93	49.98
		B	68.10	30.88	31.17	49.26
Qwen3.7-Max	API	A	68.83	31.33	34.87	50.54
		B	68.31	32.50	33.85	50.42
Qwen3.8-Max	API	A	68.84	31.54	34.53	50.54
		B	68.35	32.71	33.83	50.50
Transcript-based systems (Qwen3-ASR-Flash-Filetrans)						
DeepSeek-V4-Pro	API	A	61.77	35.21	26.67	46.48
		B	61.68	34.21	29.64	46.77
GLM-5.2	API	A	69.36	32.79	24.19	48.98
		B	69.92	32.50	23.47	49.02
Qwen3.5-Plus	API	A	67.28	31.13	33.50	49.42
		B	67.07	31.33	30.44	48.74
Qwen3.5-Omni-Plus	API	A	70.06	30.54	27.89	49.43
		B	70.07	30.58	27.15	49.30
Qwen3.8-Max	API	A	68.64	31.25	34.68	50.38
		B	67.92	32.62	33.24	50.13
Qwen3.7-Max	API	A	68.33	31.50	34.62	50.29
		B	67.93	33.00	33.57	50.32
Direct-audio systems						
Kimi-Audio-7B-Instruct	Local	A	29.46	14.12	15.06	21.87
		B	29.54	14.00	15.62	21.99
Baichuan-Omni-1.5	Local	A	28.94	15.96	19.99	23.20
		B	28.56	16.67	19.55	23.13
Qwen2.5-Omni-3B	Local	A	28.50	15.67	26.62	24.30
		B	27.84	14.37	25.12	23.27
MiniCPM-o-4.5	Local	A	40.41	17.75	25.66	30.58
		B	40.73	15.88	24.62	29.96
Qwen2.5-Omni-7B	Local	A	35.08	23.75	32.10	31.09
		B	35.54	22.71	32.83	31.16
MiMo-Audio-7B-Instruct	Local	A	42.67	22.79	20.33	32.06
		B	42.30	22.62	24.23	32.65

Continued on next page

System	Interface	Route	D	I	A	Overall
Direct-audio systems (continued)						
MOSS-Audio-8B-Instruct	Local	A	43.03	19.88	36.45	34.78
		B	42.95	21.25	34.78	34.79
Qwen3-Omni-30B-A3B-Instruct	Local	A	49.71	30.92	42.06	42.52
		B	49.60	31.67	44.40	43.18
Qwen3.5-Omni-Flash	API	A	55.70	34.00	43.73	46.74
		B	55.60	34.50	44.48	47.00
Qwen3.5-Omni-Plus	API	A	69.69	45.46	58.76	60.20
		B	69.62	47.17	57.70	60.45
Direct-audio systems (continued)						
Qwen2.5-Omni-3B-SFT	Local	A	67.54	78.46	72.81	71.89
		B	68.31	79.46	71.92	72.38
Qwen2.5-Omni-7B-SFT	Local	A	71.69	85.50	76.37	76.78
		B	71.86	86.04	75.96	76.93
Qwen3-Omni-30B-A3B-Instruct-SFT	Local	A	76.58	82.08	80.50	79.04
		B	75.47	82.75	80.75	78.74

Table 16: Route consistency across the 43 reported uniformly judged configurations. Signed differences are Route A minus Route B; absolute differences and maxima are in score points. These are descriptive configuration-level summaries, not confidence intervals or equivalence tests.

Metric	Mean $A - B$	Mean $\lvert A - B \rvert$	Max. $\lvert A - B \rvert$	Spearman ρ
D	+0.00	0.35	1.10	0.9725
I	-0.09	0.81	1.88	0.9318
A	+0.44	1.28	3.90	0.9622
Overall	+0.07	0.30	1.12	0.9829

Figure 4 compares Overall across routes, and Table 16 reports dimension statistics for the recorded items. The Overall mean absolute difference is 0.30 points with Spearman correlation 0.983; restricting both routes to exactly matched items gives the same rounded values. Dimension changes can be larger: the largest absolute A change is 3.90 points for MiMo-Audio-7B-Instruct. High rank agreement therefore does not imply that each dimension is numerically invariant.

Table 17: Direct-minus-transcript gaps for Qwen3.5-Omni-Plus, with Qwen3-ASR-Flash as the transcript frontend. Brackets give pointwise 95% paired-source intervals.

Route	D gap [95% CI]	I gap [95% CI]	A gap [95% CI]
A	-0.51 [-1.44, 0.38]	+14.92 [13.17, 16.62]	+30.67 [27.63, 33.77]
B	-0.84 [-1.75, 0.11]	+16.46 [14.46, 18.46]	+29.97 [26.64, 33.19]

Table 17 preserves the central pathway profile on both routes: D intervals include zero, while I/A gaps are positive. At the condition level, the I gap is -4.00 at L0 on each route and positive at L1–L5; every A gap is positive. Pooled cross-route differences in these pathway gaps have intervals spanning zero. However, the L4 A gap is 40.05 on Route A and 32.01 on Route B: the paired difference-in-differences is 8.03 $[3.02, 13.11]$. No D pathway-by-route interaction excludes zero at the pointwise 95% level. These pointwise intervals are unadjusted for multiple comparisons and identify local sensitivity. No equivalence margin was prespecified, and both implementations share plans and reference voices. The comparison supports consistency of the principal reported findings within this synthesis design.

Table 18 provides all condition-level pathway gaps and their cross-route differences, including the L4 acoustic–temporal contrast discussed above.

Table 18: Condition-level answering-pathway gaps by synthesis route for Qwen3.5-Omni-Plus. Each gap is direct audio minus lexical Qwen3-ASR-Flash input; the interaction is the Route A gap minus the Route B gap. Values are QA-average score points on the recorded eligibility sets, with paired, pointwise 95% source-cluster intervals. Differences use unrounded scores; intervals are unadjusted for multiple comparisons.

Condition	Dimension	Route A gap [95% CI]	Route B gap [95% CI]	Interaction [95% CI]
L0	D	+0.41 [-1.26, 2.03]	-0.41 [-2.02, 1.17]	+0.83 [-1.24, 2.89]
	I	-4.00 [-7.25, -1.00]	-4.00 [-7.25, -0.75]	+0.00 [-3.00, 2.75]
	A	+29.09 [23.95, 33.99]	+29.66 [23.94, 35.15]	-0.57 [-6.03, 5.01]
L1	D	-1.73 [-3.38, -0.15]	-0.38 [-2.20, 1.47]	-1.36 [-3.75, 1.03]
	I	+15.50 [12.75, 18.25]	+18.00 [15.00, 21.00]	-2.50 [-5.75, 0.75]
	A	+24.67 [20.00, 29.33]	+27.00 [22.00, 32.00]	-2.33 [-8.33, 3.67]
L2	D	-0.75 [-2.47, 0.87]	-0.23 [-1.92, 1.56]	-0.53 [-2.48, 1.41]
	I	+16.25 [12.00, 20.75]	+20.50 [16.25, 24.75]	-4.25 [-9.00, 0.25]
	A	+29.79 [24.84, 34.73]	+31.14 [25.94, 36.31]	-1.35 [-7.01, 4.35]
L3	D	+0.19 [-1.29, 1.66]	-1.02 [-2.73, 0.82]	+1.20 [-0.76, 3.19]
	I	+29.50 [24.50, 34.25]	+30.50 [25.00, 36.25]	-1.00 [-6.25, 4.25]
	A	+45.43 [41.33, 49.42]	+42.55 [37.82, 47.11]	+2.88 [-1.77, 7.58]
L4	D	-0.04 [-1.38, 1.33]	-1.20 [-2.93, 0.50]	+1.17 [-0.80, 3.12]
	I	+18.50 [13.75, 23.75]	+16.75 [12.00, 22.00]	+1.75 [-2.00, 5.75]
	A	+40.05 [34.95, 45.26]	+32.01 [27.14, 37.02]	+8.03 [3.02, 13.11]
L5	D	-1.13 [-2.80, 0.56]	-1.81 [-3.34, -0.23]	+0.68 [-1.68, 3.05]
	I	+13.75 [9.75, 17.75]	+17.00 [12.75, 21.25]	-3.25 [-7.25, 0.50]
	A	+14.51 [9.46, 19.50]	+17.36 [12.31, 22.32]	-2.85 [-8.41, 2.68]

C.3.2 JUDGE SENSITIVITY

The sensitivity analysis keeps candidate answers frozen and replaces Qwen3.7-Max with DeepSeek-V4-Pro for the subjective *dialogue content* tasks (D-1/D-3/D-5/D-6/D-7). It covers 5,784 subjective positions per configuration and 248,712 across the 43 reported test configurations.

Table 19 reports dialogue-content and Overall scores under both judges for all 43 configurations. Dialogue-content scores include both choice-based and subjective tasks; only the latter are regraded. Interaction-structure and acoustic–temporal scores are unchanged and are therefore not repeated. All differences use the same evaluated items and are computed before rounding.

Table 19: Judge sensitivity for all 43 reported test configurations with candidate answers held fixed. Dialogue content (D) and Overall are QA-average scores; Qwen and DeepSeek denote Qwen3.7-Max and DeepSeek-V4-Pro, respectively. Δ is DeepSeek minus Qwen, computed before rounding. D and Overall are stacked for each system.

System	Interface	Metric	Qwen	DeepSeek	Δ
Transcript-based systems (Fun-ASR)					
DeepSeek-V4-Pro	API	D	62.48	62.07	-0.41
		Overall	46.87	46.67	-0.20
GLM-5.2	API	D	68.94	67.09	-1.85
		Overall	48.64	47.73	-0.91
Qwen3.5-Plus	API	D	66.96	66.68	-0.27
		Overall	49.02	48.89	-0.13

Continued on next page

System	Interface	Metric	Qwen	DeepSeek	Δ
Transcript-based systems (Fun-ASR) (continued)					
Qwen3.5-Omni-Plus	API	D	69.56	67.01	−2.55
		Overall	49.22	47.97	−1.25
Qwen3.7-Max	API	D	67.78	67.78	+0.00
		Overall	50.02	50.03	+0.00
Qwen3.8-Max	API	D	68.14	67.81	−0.33
		Overall	50.20	50.04	−0.16
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash)					
DeepSeek-V4-Pro	API	D	61.47	61.19	−0.28
		Overall	46.15	46.02	−0.14
GLM-5.2	API	D	69.32	67.39	−1.92
		Overall	48.83	47.89	−0.95
Qwen3.5-Plus	API	D	67.21	67.08	−0.13
		Overall	48.88	48.81	−0.06
Qwen3.5-Omni-Plus	API	D	69.74	66.95	−2.79
		Overall	49.29	47.92	−1.37
Qwen3.7-Max	API	D	68.20	68.08	−0.13
		Overall	50.40	50.33	−0.06
Qwen3.8-Max	API	D	68.39	68.22	−0.17
		Overall	50.40	50.32	−0.08
Transcript-based systems (Qwen-Audio-3.0-ASR-Flash-Filetrans)					
DeepSeek-V4-Pro	API	D	61.70	61.48	−0.23
		Overall	46.54	46.43	−0.11
Qwen3.5-Plus	API	D	67.02	66.86	−0.16
		Overall	48.81	48.73	−0.08
GLM-5.2	API	D	69.02	67.42	−1.61
		Overall	48.82	48.03	−0.79
Qwen3.5-Omni-Plus	API	D	69.46	67.26	−2.21
		Overall	49.07	47.99	−1.09
Qwen3.8-Max	API	D	67.91	68.03	+0.12
		Overall	50.04	50.10	+0.06
Qwen3.7-Max	API	D	67.92	67.83	−0.09
		Overall	50.09	50.04	−0.05
Transcript-based systems (Qwen3-ASR-Flash)					
DeepSeek-V4-Pro	API	D	61.82	61.51	−0.30
		Overall	46.67	46.52	−0.15
GLM-5.2	API	D	70.05	68.19	−1.86
		Overall	49.41	48.50	−0.92
Transcript-based systems (Qwen3-ASR-Flash) (continued)					
Qwen3.5-Omni-Plus	API	D	70.33	67.84	−2.49
		Overall	49.60	48.37	−1.23
Qwen3.5-Plus	API	D	68.37	67.95	−0.42
		Overall	49.62	49.42	−0.21
Qwen3.7-Max	API	D	68.57	68.38	−0.19
		Overall	50.48	50.38	−0.09
Qwen3.8-Max	API	D	68.60	68.60	+0.00
		Overall	50.52	50.52	+0.00

Continued on next page

System	Interface	Metric	Qwen	DeepSeek	Δ
Transcript-based systems (Qwen3-ASR-Flash-Filetrans)					
DeepSeek-V4-Pro	API	D	61.73	61.36	−0.36
		Overall	46.63	46.45	−0.18
GLM-5.2	API	D	69.64	67.92	−1.72
		Overall	49.00	48.15	−0.84
Qwen3.5-Plus	API	D	67.18	66.85	−0.33
		Overall	49.08	48.92	−0.16
Qwen3.5-Omni-Plus	API	D	70.06	67.52	−2.54
		Overall	49.37	48.11	−1.25
Qwen3.8-Max	API	D	68.28	68.13	−0.15
		Overall	50.26	50.18	−0.07
Qwen3.7-Max	API	D	68.13	67.85	−0.28
		Overall	50.30	50.17	−0.14
Direct-audio systems					
Kimi-Audio-7B-Instruct	Local	D	29.50	18.69	−10.81
		Overall	21.93	16.61	−5.32
Baichuan-Omni-1.5	Local	D	28.75	19.08	−9.67
		Overall	23.17	18.40	−4.76
Qwen2.5-Omni-3B	Local	D	28.17	15.71	−12.47
		Overall	23.79	17.65	−6.14
MiniCPM-o-4.5	Local	D	40.57	36.67	−3.90
		Overall	30.27	28.35	−1.92
Qwen2.5-Omni-7B	Local	D	35.31	27.79	−7.52
		Overall	31.13	27.42	−3.70
MiMo-Audio-7B-Instruct	Local	D	42.49	36.77	−5.72
		Overall	32.35	29.54	−2.81
Direct-audio systems (continued)					
MOSS-Audio-8B-Instruct	Local	D	42.99	36.47	−6.53
		Overall	34.78	31.57	−3.21
Qwen3-Omni-30B-A3B-Instruct	Local	D	49.65	47.76	−1.90
		Overall	42.85	41.92	−0.93
Qwen3.5-Omni-Flash	API	D	55.65	52.86	−2.79
		Overall	46.87	45.49	−1.37
Qwen3.5-Omni-Plus	API	D	69.66	67.31	−2.35
		Overall	60.32	59.16	−1.16
Direct-audio systems (continued)					
Qwen2.5-Omni-3B-SFT	Local	D	67.92	68.89	+0.96
		Overall	72.13	72.61	+0.47
Qwen2.5-Omni-7B-SFT	Local	D	71.77	72.71	+0.94
		Overall	76.85	77.31	+0.46
Qwen3-Omni-30B-A3B-Instruct-SFT	Local	D	76.02	76.69	+0.67
		Overall	78.89	79.22	+0.33

Across all 248,712 subjective positions, including zero-score failures, DeepSeek-V4-Pro-minus-Qwen3.7-Max mean shifts range from −17.17 to +1.33 points: 38 configurations decrease and six increase. Pooled mean absolute item disagreement is 9.62 points, with 66.97% exact score agreement. Configuration-level subjective-score rankings have Spearman correlation 0.870, with a maximum absolute rank change of 14. These quantities concern the subjective subset; shared sources and configurations make the observations dependent. Figure 4 compares recomputed Overall scores while holding deterministic components fixed, yielding a higher rank correlation of 0.968 and a mean Overall decrease of 1.00 point.

The primary Qwen records contain 522 D-6 grading-contract failures: 213 phase-audit count mismatches and 309 invalid judge-JSON/content-contract outcomes. These are distinct from candidate

generation failures. DeepSeek successfully grades all 522, but only 67 receive positive scores; the other 455 remain zero. Failure recovery changes Overall for only three configurations, by at most 0.065 points. Removing the 522 grading failures and one subjective generation failure leaves 253,973 jointly successful positions. On this common subset, configuration-level mean shifts range from -18.03 to $+1.33$ points, rank correlation remains 0.870, and the maximum rank change remains 14. Pooled exact agreement is 66.93%, and mean absolute item disagreement is 9.63 points. Thus the observed subjective-score sensitivity is not explained by failure recovery. The common-subset audit is supplementary and does not replace the all-item primary scores.

Table 20 summarizes these subjective-answer comparisons for all eligible positions and for the jointly successful subset, keeping them distinct from the dimension-level and Overall results above.

Table 20: Subjective-answer judge sensitivity across the same 44 test configurations. The all-eligible analysis retains grading and generation failures at zero; the jointly scored analysis retains only identical QA positions successfully scored by both judges. Scores and differences are in points; Δ is DeepSeek-V4-Pro minus Qwen3.7-Max. Configuration statistics weight configurations equally; pooled statistics weight QA positions equally.

Measure	All eligible subjective QA	Jointly scored subjective QA
QA positions	254,496	253,973
Pooled Qwen3.7-Max QA score	57.62	57.74
Pooled DeepSeek-V4-Pro QA score	54.83	54.94
Mean configuration Δ	-2.79	-2.83
Configuration Δ range	$[-17.17, +1.33]$	$[-18.03, +1.33]$
Configuration-rank Spearman ρ	0.870	0.870
Maximum absolute rank change	14	14
Pooled exact score agreement (%)	66.97	66.93
Pooled item mean absolute difference	9.62	9.63

C.4 ASR AND STRUCTURAL-EVIDENCE DIAGNOSTICS

Table 2 compares ASR models with each answerer held fixed. The complete scores for all five ASR models and six answerers appear in Table 10, with task-level breakdowns in Tables 11 and 12. The following interventions complement that comparison by varying the information supplied to the answerer while controlling the underlying transcript.

C.4.1 SCOPE AND SEPARATION FROM ORDINARY EVALUATION

The diagnostic experiments test answerers' use of supplied evidence. The CORE subset contains scene recognition (A-2; 1,200 QA instances) and temporal localization (A-3; 1,017 QA instances), totaling 2,217 QA instances across both synthesis routes. CORE covers these two tasks within *acoustic–temporal organization* (A), rather than the entire dimension. All scores are deterministic. DeepSeek-V4-Pro, GLM-5.2, Qwen3.7-Max, and Qwen3.8-Max are evaluated on the four baseline contrasts in Table 3 and the complete visibility and field-quality factorials below. These interventions are reported separately from the main system comparison.

Scores are QA-item averages: every eligible QA contributes equally, and the score is 100 times the sum of recorded item scores divided by the eligible item count. The recorded eligibility sets and failure scores are retained. Where reported, pointwise 95% intervals use 10,000 paired source-cluster bootstrap draws with seed 1024. Both routes and matched intervention cells remain together within each sampled source; each draw divides its resampled score sum by its resampled QA count. Differences are computed from unrounded scores.

An ASR trace is produced by Qwen-Audio-3.0-ASR-Flash-Filetrans. A Gold trace is compiled from realized synthesis records and contains observed speech text, anonymous speaker identities, and timing. "Gold" denotes this reference-derived experimental trace, which does not fully represent every background sound. Mainline-only text, observed speech, ASR segmentation, and reference-defined units are distinct representations.

No standard evaluated system receives Gold fields. Replacing a field with its reference value is a privileged researcher intervention that estimates the usefulness of more accurate evidence under specified conditions. It is not an implementable correction procedure at deployment.

C.4.2 Baseline contrasts reported in the main text

Table 3 reports four simple differences with two distinct baselines. For availability, adding timestamps or speaker labels is compared with words-only input on fixed native ASR units. For correctness, all fields are visible on fixed reference-defined units; either timestamps or speaker labels are replaced by reference values while ASR words and the other structural field remain fixed. Both panels use Qwen-Audio-3.0-ASR-Flash-Filetrans outputs. Their different unitizations and baselines preclude adding effects across panels.

Table 21: Absolute CORE scores for the six conditions underlying the four main-text structural-evidence contrasts. Scores average the same 2,217 QA instances per condition.

Evidence condition	DeepSeek-V4-Pro	GLM-5.2	Qwen3.7-Max	Qwen3.8-Max
Availability: fixed ASR words and native ASR units				
Words only	20.30	19.18	20.74	20.70
Words and timestamps	38.96	37.84	37.96	37.49
Words and speaker labels	20.03	19.28	19.66	21.24
Correctness: all fields visible, fixed reference-defined units				
ASR values for all fields	37.77	36.74	35.80	37.09
Reference timestamps	42.10	40.15	40.23	41.25
Reference speaker labels	43.00	41.78	44.05	46.70

Table 22: Qwen3.8-Max structural-evidence results by task and on CORE. Changes are relative to the baseline within each panel; all entries are QA-average score points.

Evidence condition	A-2	A-3	CORE	Δ CORE
Availability: relative to words-only input				
Words only	38.17	0.09	20.70	–
Words and timestamps	37.92	36.98	37.49	+16.78
Words and speaker labels	39.17	0.09	21.24	+0.54
Correctness: relative to ASR-provided values				
ASR values for all fields	38.58	35.32	37.09	–
Reference timestamps	38.83	44.11	41.25	+4.17
Reference speaker labels	46.92	46.45	46.70	+9.62

For the six baseline conditions, Qwen3.8-Max gains $+16.78$ CORE points from adding timestamps, $+0.54$ from adding speaker labels, $+4.17$ from using reference timestamps, and $+9.62$ from using reference speaker labels (Tables 21 and 22). The timestamp-availability gain is concentrated in temporal localization. The complete factorials and complementary controls below extend these baseline contrasts.

C.4.3 Visibility: a two-field factorial on a fixed trace

Let Y_{st} be the QA-item average score when speaker visibility is $s \in \{0, 1\}$ and time visibility is $t \in \{0, 1\}$. Words, trace segmentation, order, question, and answer options remain fixed within each ASR or Gold source. Speaker identifiers are anonymous; hiding them removes that field rather than replacing it with a new diarization. The speaker and time effects are

$$\Delta_S = \tfrac{1}{2}[(Y_{10} - Y_{00}) + (Y_{11} - Y_{01})], \tag{5}$$

$$\Delta_T = \tfrac{1}{2}[(Y_{01} - Y_{00}) + (Y_{11} - Y_{10})]. \tag{6}$$

The difference-in-differences measures interaction, not either field's main effect. All four answerers have all four cells on both ASR and Gold traces. Table 23 reports their deterministic scene-recognition

and temporal-localization scores. These factorial effects average over the other field, whereas the main-text availability contrasts compare a single added field with words-only input.

Table 23: Absolute QA-item average scores in the four visibility cells. S/T indicate speaker/time visibility. ASR and Gold differ in underlying evidence and unitization, so their raw difference does not isolate one type of recognition error.

Answerer	Trace	Task	S0T0	S1T0	S0T1	S1T1
DeepSeek-V4-Pro	ASR	A-2	37.50	37.00	37.75	38.17
DeepSeek-V4-Pro	ASR	A-3	0.00	0.01	40.39	41.29
DeepSeek-V4-Pro	Gold	A-2	37.83	39.67	37.92	39.67
DeepSeek-V4-Pro	Gold	A-3	0.02	0.02	46.11	57.41
GLM-5.2	ASR	A-2	35.25	35.50	35.58	35.00
GLM-5.2	ASR	A-3	0.23	0.14	40.51	40.63
GLM-5.2	Gold	A-2	35.42	40.00	36.42	39.58
GLM-5.2	Gold	A-3	0.12	0.09	45.14	57.17
Qwen3.7-Max	ASR	A-2	38.25	36.25	37.25	35.75
Qwen3.7-Max	ASR	A-3	0.08	0.09	38.79	39.29
Qwen3.7-Max	Gold	A-2	37.58	41.75	37.17	42.33
Qwen3.7-Max	Gold	A-3	0.05	0.05	44.13	57.88
Qwen3.8-Max	ASR	A-2	38.17	39.17	37.92	38.08
Qwen3.8-Max	ASR	A-3	0.09	0.09	36.98	38.77
Qwen3.8-Max	Gold	A-2	38.33	44.17	38.00	45.00
Qwen3.8-Max	Gold	A-3	0.06	0.09	42.24	56.60

Table 23 reports the absolute task scores; Table 24 completes the CORE comparison with both field effects and their interaction for every answerer and trace. Table 25 further gives selected task-level effects and source-paired intervals. Across the four answerers, the ASR-trace CORE time effect is 16.97–19.11 points, whereas the speaker effect is -0.83 to $+0.73$. The corresponding Gold effects are 22.77–23.76 for time and 3.56–6.77 for speaker. This contrast motivates a separate test of field quality, but is not by itself a decomposition: the two trace sources also differ in words and units.

Table 24: CORE visibility effects on ASR and Gold traces. Speaker and time effects average over the other field; interaction is $Y_{11} - Y_{10} - Y_{01} + Y_{00}$. Entries give estimates and pointwise paired-source 95% intervals in score points.

Answerer	Trace	Speaker	Time	Interaction
		$+0.19$	$+19.11$	$+0.90$
DeepSeek-V4-Pro	ASR	$[-0.48, +0.86]$	$[+18.21, +20.02]$	$[-0.18, +1.96]$
		$+3.56$	$+23.76$	$+5.14$
DeepSeek-V4-Pro	Gold	$[+2.90, +4.21]$	$[+23.04, +24.50]$	$[+4.01, +6.29]$
		-0.08	$+18.48$	-0.36
GLM-5.2	ASR	$[-0.88, +0.68]$	$[+17.61, +19.35]$	$[-1.82, +1.14]$
		$+4.85$	$+23.58$	$+4.76$
GLM-5.2	Gold	$[+4.01, +5.70]$	$[+22.77, +24.38]$	$[+3.50, +6.02]$
		-0.83	$+17.46$	$+0.50$
Qwen3.7-Max	ASR	$[-1.66, -0.00]$	$[+16.64, +18.29]$	$[-0.63, +1.65]$
		$+5.68$	$+23.42$	$+6.85$
Qwen3.7-Max	Gold	$[+4.89, +6.51]$	$[+22.69, +24.11]$	$[+5.73, +7.96]$
		$+0.73$	$+16.97$	$+0.37$
Qwen3.8-Max	ASR	$[-0.15, +1.60]$	$[+16.03, +17.88]$	$[-1.05, +1.74]$
		$+6.77$	$+22.77$	$+7.21$
Qwen3.8-Max	Gold	$[+5.76, +7.77]$	$[+21.93, +23.60]$	$[+5.42, +9.06]$

Table 25: Selected fixed-trace visibility effects (QA-item average score points). Each effect averages over the other field; all intervals are pointwise 95% source-paired intervals.

Answerer	Task	Speaker	95% CI	Time (95% CI)
DeepSeek-V4-Pro	A-2	-0.04	$[-0.96, +0.92]$	$+0.71$ $[-0.33, +1.71]$
DeepSeek-V4-Pro	A-3	$+0.45$	$[-0.30, +1.20]$	$+40.83$ $[+39.45, +42.27]$
DeepSeek-V4-Pro	CORE	$+0.19$	$[-0.48, +0.86]$	$+19.11$ $[+18.21, +20.02]$
GLM-5.2	A-2	-0.17	$[-1.46, +1.08]$	-0.08 $[-1.12, +0.96]$
GLM-5.2	A-3	$+0.01$	$[-0.85, +0.91]$	$+40.39$ $[+39.11, +41.67]$
GLM-5.2	CORE	-0.08	$[-0.88, +0.68]$	$+18.48$ $[+17.61, +19.35]$
Qwen3.7-Max	A-2	-1.75	$[-3.12, -0.46]$	-0.75 $[-1.62, +0.12]$
Qwen3.7-Max	A-3	$+0.25$	$[-0.50, +0.98]$	$+38.95$ $[+37.41, +40.45]$
Qwen3.7-Max	CORE	-0.83	$[-1.66, -0.00]$	$+17.46$ $[+16.64, +18.29]$
Qwen3.8-Max	A-2	$+0.58$	$[-0.92, +2.00]$	-0.67 $[-1.83, +0.50]$
Qwen3.8-Max	A-3	$+0.89$	$[+0.11, +1.68]$	$+37.78$ $[+36.08, +39.46]$
Qwen3.8-Max	CORE	$+0.73$	$[-0.15, +1.60]$	$+16.97$ $[+16.03, +17.88]$

C.4.4 Quality: an eight-cell control at fixed speech units

Table 26 reports the complete eight-cell field-quality design for all four answerers. It fixes reference-defined observed-speech units, their order, and the response contract, then independently selects ASR or reference values for text, time, and speaker. All fields remain visible. Each main effect averages the four matched substitutions over the other two fields; the all-fields contrast compares the two diagonal cells. Main effects need not sum to the all-fields difference because interactions may remain. The main-text correctness contrasts instead replace one field at the all-ASR-value baseline while keeping ASR words fixed.

Table 26: Complete field-quality factorial on CORE for all four answerers. All fields are visible on fixed reference-defined units; each field takes ASR or reference values. Scores are QA averages.

Field values			CORE QA-average score			
Words	Time	Speaker	DeepSeek-V4-Pro	GLM-5.2	Qwen3.7-Max	Qwen3.8-Max
ASR	ASR	ASR	37.77	36.74	35.80	37.09
ASR	ASR	Reference	43.00	41.78	44.05	46.70
ASR	Reference	ASR	42.10	40.15	40.23	41.25
ASR	Reference	Reference	48.26	47.00	49.82	51.28
Reference	ASR	ASR	38.04	36.48	35.63	37.26
Reference	ASR	Reference	42.25	42.07	44.33	46.22
Reference	Reference	ASR	42.52	40.88	40.59	41.84
Reference	Reference	Reference	47.75	47.71	49.04	50.32

To construct these cells, minimum-edit alignment projects ASR text into the fixed units while exactly reconstructing the target lexical string. ASR timing is the outer envelope of overlapping ASR segments, with nearest-segment fallback when no overlap exists. ASR speaker clusters are first aligned to reference identities by maximum-overlap Hungarian matching; the duration-dominant aligned cluster supplies each unit's speaker value. This label permutation removes arbitrary ID naming and itself uses reference information. Consequently, even the all-ASR-value cell is conditional on reference-defined units and alignment, not a native ASR deployment baseline.

Projection checks cover 1,200 recordings and 60,534 speech units. There are 52 no-overlap timing fallbacks and 25 empty projected text units; target lexical reconstruction is exact for every recording. The all-Gold contexts reproduce their native reference contexts exactly. These checks validate the constructed contrasts. Mean boundary error (588.03 ms) and speaker-unit agreement (91.54%) are measured relative to synthesis records and do not constitute independently verified acoustic error rates.

Table 27: Field-quality effects on CORE in QA-item average score points, conditional on fixed reference-defined units. Positive values favor reference over ASR values. "All fields" is the joint diagonal contrast, not the sum of main effects. Intervals are pointwise source-paired 95% intervals.

Answerer	Replaced field	CORE effect	95% CI
DeepSeek-V4-Pro	Text	-0.14	$[-0.63, +0.34]$
DeepSeek-V4-Pro	Time	$+4.89$	$[+4.21, +5.57]$
DeepSeek-V4-Pro	Speaker	$+5.21$	$[+4.48, +5.94]$
DeepSeek-V4-Pro	All fields	$+9.98$	$[+8.90, +11.07]$
GLM-5.2	Text	$+0.37$	$[-0.26, +1.01]$
GLM-5.2	Time	$+4.67$	$[+3.79, +5.56]$
GLM-5.2	Speaker	$+6.08$	$[+5.19, +6.97]$
GLM-5.2	All fields	$+10.97$	$[+9.42, +12.51]$
Qwen3.7-Max	Text	-0.08	$[-0.50, +0.34]$
Qwen3.7-Max	Time	$+4.97$	$[+4.21, +5.76]$
Qwen3.7-Max	Speaker	$+8.75$	$[+7.71, +9.77]$
Qwen3.7-Max	All fields	$+13.24$	$[+11.81, +14.70]$
Qwen3.8-Max	Text	-0.17	$[-0.76, +0.41]$
Qwen3.8-Max	Time	$+4.35$	$[+3.57, +5.14]$
Qwen3.8-Max	Speaker	$+9.27$	$[+8.20, +10.37]$
Qwen3.8-Max	All fields	$+13.24$	$[+11.63, +14.84]$

Table 27 shows gains of 4.35–4.97 points from time quality and 5.21–9.27 from speaker quality across the four answerers. Text substitutions range from -0.17 to $+0.37$ points, with all intervals including zero; the all-fields contrast ranges from 9.98 to 13.24 points. The conclusion is specific to this CORE slice, source frontend, and projection procedure. It neither establishes that lexical errors are harmless nor identifies a pure segmentation effect. Comparisons against native-unit runs additionally contain unitization and independent generation differences; we do not use their residual as an error decomposition.

C.4.5 Correctness controls and lexical robustness

Timestamp corruption. The Gold timing control changes segment start/end values while freezing words, anonymous speaker partitions, questions, and options. Correct and shifted contexts retain the same evidence schema. The recorded corruption uses seed 1024. Table 28 reports its paired A-3 results. The large loss under incorrect values strengthens the distinction between having a time field and having usable time evidence; it does not measure acoustic time estimation.

Table 28: Correct versus shifted reference timestamps on A-3. Scores are QA-item averages over matched items, with differences in score points; intervals are pointwise paired-source 95% intervals.

Answerer	Correct	Shifted	Difference	95% CI
DeepSeek-V4-Pro	57.41	5.86	$+51.55$	$[+49.60, +53.35]$
GLM-5.2	57.17	6.07	$+51.10$	$[+49.36, +52.75]$
Qwen3.7-Max	57.88	5.77	$+52.11$	$[+50.20, +53.90]$
Qwen3.8-Max	56.60	6.02	$+50.58$	$[+48.53, +52.47]$

Complementary speaker substitution. A complementary control holds native ASR words, order, and times fixed and replaces only speaker assignments with reference-supported assignments. Table 29 reports the original and repaired scores: scene recognition increases by 1.50 points for DeepSeek-V4-Pro, 3.08 for GLM-5.2, 6.25 for Qwen3.7-Max, and 8.00 for Qwen3.8-Max. This is an oracle diagnostic. Its native ASR units differ from the fixed-reference-unit factorial, so the effects are not interchangeable or additive.

Table 29: Paired speaker-substitution control on scene recognition (A-2; 1,200 QA instances). Native ASR words, order, and times remain fixed; repaired assignments use reference support. These original/repaired arms belong to this control, not to the separate visibility factorial. Scores are QA averages; differences are repaired minus original with pointwise paired-source 95% intervals.

Answerer	Original	Repaired	Difference	95% CI
DeepSeek-V4-Pro	37.92	39.42	$+1.50$	$[0.33, 2.75]$
GLM-5.2	35.08	38.17	$+3.08$	$[1.42, 4.83]$
Qwen3.7-Max	35.42	41.67	$+6.25$	$[4.67, 7.83]$
Qwen3.8-Max	38.08	46.08	$+8.00$	$[6.00, 10.17]$

Five lexical sources on one structural carrier. The five-ASR follow-up fixes the segment boundaries, speaker values, and time values from Qwen-Audio-3.0-ASR-Flash-Filetrans and projects five ASR lexical strings onto that carrier. The sources are Fun-ASR, Qwen-Audio-3.0-ASR-Flash, Qwen-Audio-3.0-ASR-Flash-Filetrans, Qwen3-ASR-Flash, and Qwen3-ASR-Flash-Filetrans. It then repeats the speaker/time visibility factorial. Across 6,000 projected recording–frontend pairs, every target lexical string reconstructs exactly. Thus, this experiment tests sensitivity to lexical variation at one fixed structural interface, not the native timestamp or diarization quality of five ASR systems.

Table 30: Ranges of visibility effects across five projected lexical sources on a shared structural carrier (QA-item average score points). These ranges summarize robustness across the recorded lexical variants, not confidence intervals.

Answerer	CORE speaker	CORE time	Localization time
DeepSeek-V4-Pro	-0.32 to $+0.57$	$+19.02$ to $+19.59$	$+40.43$ to $+41.03$
GLM-5.2	$+0.49$ to $+0.80$	$+17.64$ to $+18.66$	$+39.70$ to $+40.46$
Qwen3.7-Max	-1.14 to -0.61	$+17.15$ to $+17.84$	$+38.77$ to $+39.39$
Qwen3.8-Max	$+0.87$ to $+1.33$	$+16.78$ to $+17.25$	$+37.25$ to $+37.86$

The temporal effect remains positive across all five lexical sources for all four answerers (Table 30). Table 31 supplies the underlying CORE scores for every visibility cell, and Table 32 reports each lexical source's field-effect differences relative to Qwen-Audio-3.0-ASR-Flash-Filetrans. Among these 48 contrasts, four pointwise intervals exclude zero; all contrast magnitudes are below two points. These comparisons describe sensitivity to lexical variation without a multiplicity-corrected omnibus claim.

Table 31: All CORE visibility cells for five lexical sources on a shared structural carrier. Words from each listed ASR frontend are projected onto the fixed segment, speaker, and time fields of Qwen-Audio-3.0-ASR-Flash-Filetrans. S/T indicate speaker/time visibility. Scores average the same 2,217 QA instances; these are not native-frontend end-to-end comparisons.

Lexical source	S0T0	S1T0	S0T1	S1T1
DeepSeek-V4-Pro				
Fun-ASR	19.67	20.17	38.67	39.32
Qwen-Audio-3.0-ASR-Flash	20.43	20.17	39.51	39.13
Qwen-Audio-3.0-ASR-Flash-Filetrans	20.34	19.49	38.58	39.59
Qwen3-ASR-Flash	19.72	19.81	39.00	39.70
Qwen3-ASR-Flash-Filetrans	19.89	20.08	38.92	39.60

Continued on next page

Lexical source	S0T0	S1T0	S0T1	S1T1
GLM-5.2				
Fun-ASR	18.91	19.51	36.61	37.56
Qwen-Audio-3.0-ASR-Flash	18.89	18.95	36.77	37.68
Qwen-Audio-3.0-ASR-Flash-Filetrans	18.87	19.28	37.13	38.34
Qwen3-ASR-Flash	19.10	18.95	36.70	38.20
Qwen3-ASR-Flash-Filetrans	19.19	19.63	36.65	37.46
Qwen3.7-Max				
Fun-ASR	20.59	19.79	37.79	37.37
Qwen-Audio-3.0-ASR-Flash	20.74	19.83	37.82	37.06
Qwen-Audio-3.0-ASR-Flash-Filetrans	20.54	19.65	38.30	37.58
Qwen3-ASR-Flash	20.50	19.80	38.07	36.94
Qwen3-ASR-Flash-Filetrans	20.79	20.15	38.56	36.92
Qwen3.8-Max				
Fun-ASR	20.68	21.20	36.90	38.55
Qwen-Audio-3.0-ASR-Flash	20.55	20.98	37.20	38.51
Qwen-Audio-3.0-ASR-Flash-Filetrans	20.48	20.84	37.03	38.53
Qwen3-ASR-Flash	20.86	21.43	36.97	39.07
Qwen3-ASR-Flash-Filetrans	20.42	20.97	37.21	38.67

Table 32: The 48 CORE lexical-source contrasts underlying the five-ASR robustness analysis. Each entry subtracts the matching speaker effect, time effect, or speaker–time interaction under the Qwen-Audio-3.0-ASR-Flash-Filetrans lexical source from its value under the listed source. Estimates and pointwise paired-source 95% intervals are in score points; the structural carrier and eligible QA instances remain fixed.

Lexical source	Speaker effect	Time effect	Interaction
DeepSeek-V4-Pro			
Fun-ASR	$+0.49$ $[-0.23, +1.21]$	-0.09 $[-0.88, +0.66]$	-1.70 $[-2.98, -0.44]$
Qwen-Audio-3.0-ASR-Flash	-0.40 $[-1.03, +0.24]$	-0.15 $[-0.98, +0.61]$	-1.97 $[-3.44, -0.51]$
Qwen3-ASR-Flash	$+0.32$ $[-0.46, +1.12]$	$+0.42$ $[-0.34, +1.17]$	-1.24 $[-2.57, +0.09]$
Qwen3-ASR-Flash-Filetrans	$+0.36$ $[-0.33, +1.03]$	$+0.10$ $[-0.63, +0.82]$	-1.36 $[-2.73, +0.07]$
GLM-5.2			
Fun-ASR	-0.03 $[-0.90, +0.82]$	-0.78 $[-1.82, +0.27]$	-0.45 $[-2.58, +1.69]$
Qwen-Audio-3.0-ASR-Flash	-0.32 $[-1.33, +0.67]$	-0.36 $[-1.31, +0.63]$	$+0.05$ $[-2.02, +2.16]$
Qwen3-ASR-Flash	-0.13 $[-1.09, +0.84]$	-0.23 $[-1.12, +0.67]$	$+0.84$ $[-1.06, +2.73]$
Qwen3-ASR-Flash-Filetrans	-0.18 $[-1.11, +0.76]$	-1.02 $[-2.10, +0.05]$	-0.44 $[-2.41, +1.55]$

Continued on next page

Lexical source	Speaker effect	Time effect	Interaction
Qwen3.7-Max			
	$+0.20$	-0.45	$+0.22$
Fun-ASR	$[-0.41, +0.83]$	$[-1.05, +0.14]$	$[-1.03, +1.46]$
	-0.03	-0.69	-0.03
Qwen-Audio-3.0-ASR-Flash	$[-0.71, +0.67]$	$[-1.36, -0.01]$	$[-1.54, +1.42]$
	-0.11	-0.48	-0.60
Qwen3-ASR-Flash	$[-0.75, +0.52]$	$[-1.06, +0.08]$	$[-1.78, +0.57]$
	-0.34	-0.57	-1.16
Qwen3-ASR-Flash-Filetrans	$[-0.95, +0.27]$	$[-1.15, -0.00]$	$[-2.59, +0.21]$
Qwen3.8-Max			
	$+0.15$	-0.34	-0.02
Fun-ASR	$[-0.60, +0.90]$	$[-1.07, +0.41]$	$[-1.64, +1.57]$
	-0.06	-0.03	-0.26
Qwen-Audio-3.0-ASR-Flash	$[-0.99, +0.92]$	$[-0.76, +0.72]$	$[-2.14, +1.63]$
	$+0.40$	-0.24	$+0.38$
Qwen3-ASR-Flash	$[-0.48, +1.30]$	$[-0.96, +0.49]$	$[-1.39, +2.14]$
	$+0.07$	$+0.13$	-0.24
Qwen3-ASR-Flash-Filetrans	$[-0.84, +0.97]$	$[-0.68, +0.94]$	$[-1.87, +1.42]$

C.4.6 Legacy structure and audio-input attribution

Legacy versus compact representation. The legacy representation includes segment speaker/time fields, the whole transcript, and word timestamps, changing both evidence and serialization redundancy. Table 33 compares it with lexical and compact S1T1 inputs on CORE. Gains over lexical input range from 16.41 to 20.29 points; differences from compact S1T1 are much smaller and not uniformly positive. These results support the value of structural evidence without attributing the gains to verbose serialization.

Table 33: CORE scores for lexical, compact, and legacy inputs, with legacy-minus-baseline differences and pointwise paired-source 95% intervals. Compact S1T1 already exposes speaker and time.

Answerer	Lexical	Compact	Legacy	Legacy − lexical 95% CI	Legacy − compact 95% CI
DeepSeek-V4-Pro	20.60	39.60	38.98	+18.38 [17.31, 19.45]	-0.61 [-1.70, 0.46]
GLM-5.2	17.93	37.58	38.22	+20.29 [19.30, 21.26]	+0.63 [-0.55, 1.83]
Qwen3.7-Max	22.66	37.37	39.06	+16.41 [15.08, 17.69]	+1.69 [0.51, 2.88]
Qwen3.8-Max	22.72	38.40	40.33	+17.62 [16.27, 18.96]	+1.93 [0.86, 3.02]

Recording substitution at a fixed transcript. Table 34 reports complementary controls for Qwen3.5-Omni-Plus with audio available. Holding the recording and words fixed, adding speaker/time fields improves CORE by 2.48 points with ASR text and 11.39 with reference text. These joint effects differ from the text-only field effects because audio is already available. Correct reference speaker and time values also outperform corrupted values on A-2 and A-3.

Replacing unrelated audio with target audio at a fixed ASR S1T1 trace improves the average across the seven audio tasks by 12.42 points $[11.16, 13.72]$. The recording supplies useful information beyond that transcript interface, without isolating the responsible acoustic cues. Reference-field rows remain oracle diagnostics. These audio-input interventions do not apply to transcript-only configurations; absent cells are not imputed or treated as failures.

Table 34: Input-control scores for Qwen3.5-Omni-Plus. Left and Right denote the first and second conditions named in each contrast; differences include pointwise paired-source 95% intervals. Task scopes differ and effects are not additive.

Contrast	Task scope	Left	Right	Difference [95% CI]
ASR S1T1 minus ASR S0T0, fixed audio	A-2+A-3	41.68	39.20	+2.48 [1.16, 3.76]
Gold S1T1 minus Gold S0T0, fixed audio	A-2+A-3	50.69	39.31	+11.39 [10.22, 12.55]
Correct minus corrupted reference speakers	A-2	42.08	34.17	+7.92 [6.08, 9.75]
Correct minus shifted reference times	A-3	54.14	5.11	+49.03 [46.59, 51.31]
Target minus unrelated audio, fixed ASR trace	I-1–I-4 / A-1–A-3	50.68	38.25	+12.42 [11.16, 13.72]

C.5 Adaptation results

We compare three models before and after supervised fine-tuning (SFT) on the ConvoAudit training set. Training details are provided in Appendix B.1. All comparisons use the source-disjoint test set and the same evaluation protocol.

Table 35: Test performance before and after SFT. Columns report dialogue content (D), interaction structure (I), acoustic–temporal organization (A), and Overall.

	Before SFT				**After SFT**			
Model	D	I	A	Overall	D	I	A	Overall
Qwen2.5-Omni-3B	28.17	15.02	25.87	23.79	67.92	78.96	72.37	72.13
Qwen2.5-Omni-7B	35.31	23.23	32.47	31.13	71.77	85.77	76.17	76.85
Qwen3-Omni-30B-A3B-Instruct	49.65	31.29	43.23	42.85	76.02	82.42	80.62	78.89

As shown in Table 35, SFT improves all three models across all capability dimensions, yielding Overall scores of 72.13–78.89. Each adapted model exceeds the unadapted Qwen3.5-Omni-Plus baseline (60.32) under the same test scope and primary judge. This comparison excludes the Gemini configurations evaluated only on test-mini.

Table 36: SFT-minus-Base gains on test, with paired source-cluster 95% percentile intervals. Gains are reported in score points; intervals are pointwise and unadjusted for multiple comparisons.

Base/SFT pair	D gain	I gain	A gain	Overall gain
Qwen2.5-Omni-3B	+39.75 [37.49, 41.98]	+63.94 [62.29, 65.50]	+46.49 [43.85, 49.12]	+48.35 [47.02, 49.65]
Qwen2.5-Omni-7B	+36.46 [34.24, 38.66]	+62.54 [60.71, 64.38]	+43.70 [41.26, 46.14]	+45.73 [44.36, 47.14]
Qwen3-Omni-30B-A3B-Instruct	+26.37 [23.98, 28.78]	+51.12 [49.33, 52.88]	+37.39 [35.54, 39.28]	+36.04 [34.70, 37.37]

Table 36 further shows that the largest gains occur in interaction structure (51.12–63.94), followed by acoustic–temporal organization (37.39–46.49) and dialogue content (26.37–39.75). Overall gains range from 36.04 to 48.35 points, and all reported intervals have positive lower bounds. These results highlight the value of ConvoAudit for benchmark-specific model adaptation, with particularly pronounced improvements in interaction structure and acoustic–temporal organization.