SCOREQA: A SCORE-GROUNDED BENCHMARK FOR READING AND REASONING OVER SYMBOLIC SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating score-reading ability requires questions whose answers are determined by the supplied notation and explicit musical conventions. ScoreQA assesses defined Reading and Reasoning abilities on short ABC 2.1 score windows. Reading identifies events, notation attributes, and timing; Reasoning compares or transforms this information. A MusicXML-based oracle verifies answers and their evidence in ABC. Each quartet fixes one question and four choices across four valid windows supporting different answers. ARJQ describes item, paired-score, and complete-quartet correctness, together with obsolete-answer retention. Seven open-weight backbones evaluated with plain-text five-shot restricted next-token scoring achieve no complete Reasoning quartets. In a separate five-shot API evaluation with generated answer letters, five models achieve 92.8% to 99.9% *Accuracy* and 76.8% to 99.6% *Quartet*, including 72.0% to 100.0% Reasoning *Quartet*. Results are reported separately for the two prompting and inference settings. Supervised adaptation provides a supplementary use of the training split, raising Qwen3.5-9B's Reading and Reasoning *Quartet* to 84.0% and 44.0%. ScoreQA provides a verifiable test of specified score-reading and analysis abilities, for which complete correctness is an intended outcome.

1 INTRODUCTION

Symbolic scores support composition, analysis, and music education. Evaluating a model's ability to read them requires distinguishing an incorrect response from a question whose answer is not determined by the available evidence. Missing notation context or unstated musical conventions can make this distinction unclear. Conversely, general theory knowledge can suffice when a question does not require its accompanying score. An ability test should specify what the model must read or compute, supply the necessary evidence, and make the reference answer verifiable. Identifying a perfect fifth, for example, requires both an interval rule and the pitches determined by the supplied clefs, key signatures, and accidentals.

Recent resources cover theory questions, local musical calculations, and complete-score understanding (Yuan et al., 2024; Wang et al., 2025; Dai et al., 2026). MSU-Bench uses open-ended questions about complete works to test abilities from notation localization to harmony and form analysis. Its title-only control also examines the role of musical content (Dai et al., 2026). These evaluations address complementary scopes and answer formats. ScoreQA focuses on defined reading and analysis operations with answers computable from short score windows under explicit conventions. Its goal is to assess attainment of these abilities. A model that performs all tested operations correctly should receive full credit.

Symbolic scores make this dependence testable under controlled conditions. MusicXML records pitch, duration, voice, and notation state, allowing a deterministic oracle to compute unique answers and verify their preservation in model-visible ABC 2.1. ScoreQA uses this structure to organize score-grounded quartets. Four valid windows share one question and four choices, with each choice supported by a different window. The question alone cannot distinguish member labels, and correctness across the windows tests whether answers follow the current evidence. Reading covers event identification, notation states, and timing; Reasoning applies musical comparisons or transformations. ARJQ summarizes related aspects of item, pair, and complete-quartet behavior. API results show near-complete correctness for some models, alongside incomplete performance under

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.

the open-weight restricted-choice setting. Both outcomes are informative for a test with a defined correctness standard. A training split supports a supplementary adaptation study. Figure 1 summarizes the tasks, construction, and evaluation.

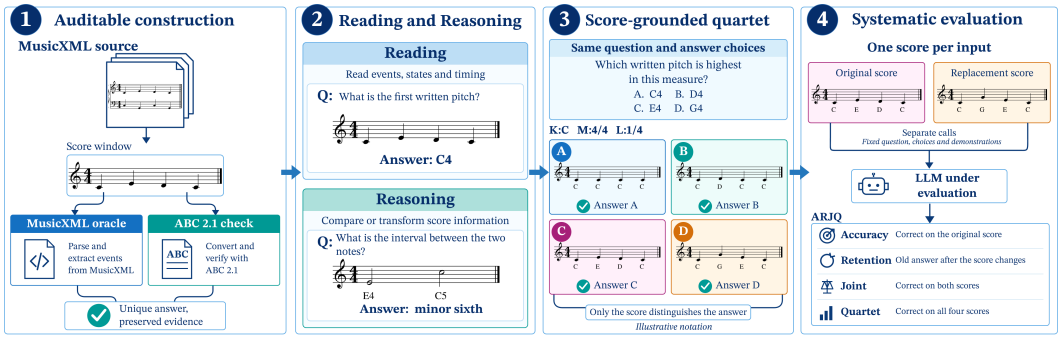

Figure 1: ScoreQA construction and evaluation. Verified Reading and Reasoning quartets support item-, pair-, and group-level ARJQ. Scores and answer badges are illustrative, not test examples or model inputs; models receive one ABC 2.1 window at a time.

Our contributions are threefold.

- **Constructing verifiable score-reading and analysis tasks.** ScoreQA covers Reading and Reasoning, uses a MusicXML oracle to determine unique answers, and checks that ABC 2.1 preserves the evidence. Quartets fix the question and choices while four valid scores support four different answers. Their construction prevents the question alone from distinguishing member labels.
- **Evaluating correctness across controlled score evidence.** ARJQ measures item correctness, obsolete-answer retention, paired correctness, and complete success across four scores for one question. It describes answers under whole-window replacements with verified answer changes.
- **Characterizing attainment under distinct evaluation settings.** Evaluation covers open-weight model families and sizes under restricted-choice scoring and five models evaluated via API using generated answers. Complete ARJQ results characterize both errors and near-complete success. Adaptation examines the supplementary utility of the training split.

2 Related Work

Sequence models have generated folk melodies (Sturm et al., 2016), modeled long-range musical structure (Huang et al., 2019), and incorporated symbolic music into language-model training and reasoning (Yuan et al., 2024; Qu et al., 2025; Deng et al., 2024; Wu et al., 2024). These studies establish that language models can process symbolic music. Generation quality or textual theory-question accuracy alone, however, does not establish that a model read a supplied score. MusicTheoryBench (MTB) combines musical knowledge and reasoning questions (Yuan et al., 2024). ABC-Eval integrates ten symbolic-music understanding and instruction-following tasks (Zhao et al., 2026). SSMR-Bench derives nine types of text and image score questions from MelodyHub and supplies rule-verified synthetic data for reinforcement learning (Wang et al., 2025). MusICA-MetaBench extracts answers programmatically from MusicXML and constructs matched questions across audio, image, and symbolic representations (Sourada et al., 2026).

MSU-Bench is a directly related score-understanding benchmark. It takes complete works as input, covers initial information, local notation, harmony, and form, and compares ABC and PDF representations (Dai et al., 2026). Music specialists supply reference answers that experts verify; three model judges score open responses by majority vote. Its Level-wise Success Rate (LSR) requires all questions through a target level to be correct for the same work, testing understanding across levels. ScoreQA instead uses verified score windows, MusicXML oracle answers, and direct answer-label matching. *Quartet* requires correct answers to the same question across four scores, testing

whether the model consistently distinguishes evidence supporting different answers. One groups levels within a work; the other groups scores under a fixed question. Table 1 summarizes these complementary evaluation designs. MSU-Bench's title-only control and MusICA-MetaBench's absent-input and noisy-input controls already examine performance when musical information is removed or disturbed (Dai et al., 2026; Sourada et al., 2026). ScoreQA complements these controls with verified answer differences under a fixed question, allowing correct updates, obsolete answers, and other errors to be distinguished.

Table 1: Evaluation designs of symbolic-score benchmarks. Question control holds questions and choices fixed across valid scores with different answers. Program verification concerns reference answers. Circles indicate partial coverage; dashes indicate unreported features.

Resource	Main focus	Scope	Answer form	Program verification	Question control	Evidence-use test
MTB	Theory QA	Mixed	MCQ	–	–	–
ABC-Eval	Syntax/sequences	Work	Mixed	○	–	–
SSMR-Bench	Local calculation	Excerpt	MCQ	✓	–	–
MSU-Bench	Multilevel QA	Full score	Open	–	–	Title only
MusICA-MetaBench	Perception/counting	Work	MCQ	✓	–	Removal or noise
ScoreQA	Reading/reasoning	Window	MCQ	✓	✓	Verified answer change

This design also relates to broader work on evidence-grounded evaluation. Multiple-choice questions support controlled response formats (Hendrycks et al., 2021), but predictions can remain sensitive to choice order and surface form (Robinson et al., 2023; Zheng et al., 2024; Sclar et al., 2024). Programmatic environments enable precise supervision (Weston et al., 2016; Johnson et al., 2017), while paired examples and minimal edits help test reliance on intended evidence (Goyal et al., 2017; Gardner et al., 2020; Sakaguchi et al., 2020). ScoreQA applies this principle to music through valid notation and explicit musical operations. Its supplementary adaptation study uses LoRA (Hu et al., 2022) and checks aggregate general-knowledge performance, following concerns about capability changes during specialization (Luo et al., 2025).

3 The ScoreQA Benchmark

3.1 Benchmark Design

ScoreQA contains Reading and Reasoning. Reading identifies events, notation attributes, and timing. Reasoning compares identified score information or applies a musical rule to it, such as computing an interval or deriving sounding pitch. Reasoning is one category contrasting these operations with information identification, rather than separate operation-specific tracks. Reading and Reasoning denote the operations required by the questions. The tasks draw on ABRSM graded theory syllabuses and official examples to cover foundational score-reading abilities and musical operations (Associated Board of the Royal Schools of Music, 2020; n.d.). Appendix A.2 gives the topic correspondence. Each item supplies an ABC 2.1 score, a question, and four choices, and asks for an answer label. Scores come from public-domain or CC0 MusicXML in PDMX (Long et al., 2025). Before window extraction, exact duplicates, arrangement families, matching non-generic composer/title identities, and transposition-invariant melodic fingerprints that include rhythm are merged into source groups. Sources are then split into training, validation, demonstration, and test groups before extracting synchronized windows of complete logical measures. Each window restores inherited clefs, key signatures, and time signatures at its left boundary. Questions about these explicit fields assess the notation state at the queried position. Windows preserve voices and empty measures and avoid cutting ties, tuplets, repeats, or cross-measure structures needed for the answer. Included windows must satisfy context-budget and voice-synchronization requirements.

Questions are generated from task-specific templates rather than written for individual pieces. For each window, the generator identifies eligible targets and fills a template with the voice, measure, and event position required by the task. For example, a written-pitch question names a voice, a measure, and an onset measured in quarter notes. A Reasoning question also specifies the comparison or transformation to apply. The MusicXML oracle is a deterministic program that computes reference

answers from source scores. It parses a quarter-note-based timeline with event onsets, durations, written pitches, sounding pitches, voices, and notation states, then applies the operation named in the question. Let s be a source score, q a question, and O the choices o_A, o_B, o_C, and o_D. The computed answer must match exactly one choice,

$$y = f(s, q), \qquad |\{i : o_i \equiv y\}| = 1. \tag{1}$$

The model does not see oracle outputs. It reads the standard ABC 2.1 representation $T(s)$ produced by a pinned version of Willem Vree's `xml2abc` and subsequent sanitization (Walshaw, 2011). Conversion removes unnecessary metadata while preserving the notation and voice information needed to answer. Representation checks require the answer recomputed from the displayed score to agree with the source oracle, and the required features to pass task-specific round-trip checks. Voices are matched by event timing and notation state rather than their identifiers in the two formats. Event identifiers, rule parameters, normalized results, and source and representation hashes make each answer traceable to musical evidence.

The oracle and model prompts use explicit reading conventions. Written pitch combines the note letter, active key signature, and applicable accidental. Sounding pitch additionally accounts for the active clef and supported voice transposition. Event counting ignores grace notes, treats a chord as one printed event, and counts tied continuations as printed events. Interval questions are restricted to unisons through octaves. Pitch targets must contain actual notes with unambiguous voice assignments. Sounding-pitch and harmonic-interval targets must fall within supported transpositions. These conventions keep programmatic verification and model responses aligned on the same musical task.

Unique answers and faithful representations ensure that questions can be answered from the score. Quartets further make the score necessary to distinguish answers. Candidate items are grouped by identical question text. Four windows from different source groups are selected with different oracle answers. The final choices come from these four computed answers, rather than from independently generated distractors. Those answers form a shared choice set, with a deterministic permutation assigning labels A through D,

$$\mathcal{G}(q, O) = \{T(s_A), T(s_B), T(s_C), T(s_D)\}, \qquad f(s_i, q) = o_i. \tag{2}$$

Each choice is correct for one member and a distractor for the others. The question and choices remain unchanged, with each label occurring once. Their content alone cannot distinguish the correct label. Each member is verified independently by the oracle, so answers under both original and replacement scores are known. This supports the item, pair, and group evaluations in Section 3.3.

For example, a Reading quartet asks for written pitch at the same position in four windows, supporting G3, G5, G#4, and G4. A Reasoning item instead compares E4 and C5 to obtain a minor sixth. Appendix A.2 traces their construction; Section 4.5 examines the model responses.

Structural checks and answer verification test two separate conditions. Structural checks confirm shared questions and choices, four distinct windows, and one occurrence of each label. All 250 test quartets satisfy these checks. Their rendered evaluation prompts are also checked for identical text and demonstration scores outside the target score. Answer verification recomputes answers from displayed ABC, checks agreement with the source MusicXML oracle and unique choice matching, and uses an independent text reader to check written pitch. All 11,165 items, including demonstrations, pass their applicable checks. Source-group and content-hash intersections across the four splits are empty. The structural checks establish that the question and choices cannot distinguish member labels; answer verification establishes that visible evidence supports a unique answer. Whether models use that evidence remains an empirical question.

Complementing these programmatic checks, two human reviewers with musical backgrounds independently answered the same 96 sampled test items. Their answers matched the reference on 89 and 95 items, respectively, giving agreement rates of 92.71% and 98.96%; the reviewers agreed with each other on 90 items, or 93.75%. This provides complementary human evidence of answerability on the reviewed sample. Appendix A.3 details the sampling and results.

3.2 Dataset Statistics

Table 2 lists operations and counts for Reading and Reasoning; Table 3 summarizes the four splits. Reading has eight notation units, each with 1,000 training, 12 validation, and 100 test items. Rea-

soning is one category of musical comparisons and transformations, such as finding the highest note, calculating an interval, or deriving sounding pitch. Training, validation, and test splits contain complete quartets with four Reading items for every Reasoning item. The demonstration split contains 45 independent items used in few-shot prompts and excluded from test results. Source separation precedes window extraction, so evaluation uses unseen sources rather than different passages from a training work. The test split draws on 442 source files in 433 source groups and contains 442 distinct ABC windows. A window can support multiple questions, so item counts exceed window counts. Appendix A.1 reports these source and representation counts for all splits.

Table 2: ScoreQA tasks and item counts, with Reasoning reported as one category.

Part	Task	Operation and answer	Train	Val.	Test
Reading	Written pitch	Written pitch at target position	1,000	12	100
	Event duration	Duration in quarter notes	1,000	12	100
	Event type	Note/rest type and duration	1,000	12	100
	Key signature	Active key signature	1,000	12	100
	Time signature	Active time signature	1,000	12	100
	Chord pitches	Written pitches within a chord	1,000	12	100
	Simultaneous pitches	Aligned written pitches across voices	1,000	12	100
	Voice clef	Clef of the specified voice	1,000	12	100
Reasoning	Musical reasoning	Compare or transform score information	2,000	24	200

Table 3: ScoreQA split sizes. Each quartet contains four items; demonstrations are independent items.

Split	Quartets	Items	Reading	Reasoning
Train	2,500	10,000	8,000	2,000
Validation	30	120	96	24
Test	250	1,000	800	200
Demonstrations	–	45	40	5

3.3 Evaluation Protocol

Evaluation uses a fixed demonstration prefix. The full-score condition presents the target window; the replacement condition presents the next quartet member in an A-to-D cycle. The two conditions use independent calls, with no previous prediction supplied to the model. Let a_i and b_i be predictions under the original and replacement conditions, ℓ_i the original correct label, and $j(i)$ the replacement member index. Indicator c_i records whether a_i matches ℓ_i; d_i records whether b_i matches $\ell_{j(i)}$; and r_i records whether b_i matches ℓ_i. For N items partitioned into K quartets with index sets $\mathcal{I}_k$, the four metrics are

$$\begin{aligned} \textit{Accuracy} &= \frac{1}{N}\sum_i c_i, \qquad \textit{Retention} = \frac{1}{N}\sum_i r_i, \\ \textit{Joint} &= \frac{1}{N}\sum_i c_i d_i, \qquad \textit{Quartet} = \frac{1}{K}\sum_k \prod_{i\in\mathcal{I}_k} c_i. \end{aligned} \tag{3}$$

Accuracy measures item accuracy with the original score. *Retention* measures agreement with the original correct label after replacement. Its reference is the original gold label, and the two predictions come from independent calls. *Joint* requires both conditions to be correct for an item; *Quartet* requires all four original quartet members to be correct. These metrics are collectively ARJQ. The replacement prompt is identical to the donor member's original prompt, not a new question. Under deterministic scoring, its prediction is therefore the donor's original prediction. *Retention* and *Joint* summarize within-quartet prediction pairs, and replacement-answer accuracy equals *Accuracy*. With sampled responses, separate calls instead provide repeated observations of the same inputs; empirical accuracies need not coincide. ARJQ describes these related correctness patterns. After score removal, identical prompts and balanced labels within quartets fix deterministic accuracy at 25%.

Retention must always be interpreted alongside correctness. Figure 2 uses quartet-cluster bootstrap intervals, with the resampling procedure specified in Appendix B.2.

4 EXPERIMENTS

4.1 SETUP

All evaluated backbones are instruction-tuned; we omit the IT and Instruct suffixes throughout. Evaluation covers Qwen3.5-9B (Qwen Team, 2026), Llama-3.1-8B (Grattafiori et al., 2024), and Gemma-2-9B (Gemma Team, 2024). Qwen3.5-0.8B, 2B, 4B, and 27B extend the comparison within one family. The open-weight baseline uses plain-text five-shot prompts and restricted next-token scoring, without chat wrappers or generated reasoning. This fixes a common evaluation interface across backbones. All 1,000 test prompts are untruncated, with demonstrations fixed within quartets. General knowledge for Qwen3.5-9B and Llama-3.1-8B is evaluated on the same 14,042-item MMLU test using subject-matched five-shot demonstrations and restricted answer-letter scoring, identically before and after adaptation. MMLU does not select checkpoints. The external zero-shot diagnostic is specified in Section 4.3.

We also evaluate ScoreQA-adapted Qwen3.5-9B and Llama-3.1-8B models, each trained once with ordinary LoRA and reported at the final one-epoch checkpoint. Training utility is examined in Section 4.4; scoring and training settings appear in Appendices C and D.

4.2 BENCHMARK FINDINGS

Table 4 compares unadapted backbones on the same held-out quartets. Figure 2 shows within-family scaling with items and scoring fixed. From Qwen3.5-0.8B to 27B, Overall *Accuracy* rises from 29.60% to 67.90% and *Quartet* from 1.20% to 33.20%. The component gap remains, with 27B Reading and Reasoning *Accuracy* of 75.25% and 38.50%.

Table 4: Held-out ARJQ before ScoreQA adaptation, in percentages. Arrows indicate preferred directions.

		Qwen3.5					Llama 3.1	Gemma 2
Scope	**Metric**	**0.8B**	**2B**	**4B**	**9B**	**27B**	**8B**	**9B**
Overall	*Accuracy* ↑	29.60	34.00	47.30	51.00	67.90	41.90	40.60
	Retention ↓	24.10	20.70	17.20	15.90	10.40	21.90	19.50
	Joint ↑	6.80	10.50	25.30	29.40	50.10	18.80	17.40
	Quartet ↑	1.20	3.20	14.00	16.40	33.20	10.40	8.80
Reading	*Accuracy* ↑	31.25	36.25	51.62	56.00	75.25	45.25	44.88
	Retention ↓	23.50	19.75	15.62	13.50	7.50	20.88	17.62
	Joint ↑	8.00	12.62	30.50	35.25	59.25	22.62	21.50
	Quartet ↑	1.50	4.00	17.50	20.50	41.50	13.00	11.00
Reasoning	*Accuracy* ↑	23.00	25.00	30.00	31.00	38.50	28.50	23.50
	Retention ↓	26.50	24.50	23.50	25.50	22.00	26.00	27.00
	Joint ↑	2.00	2.00	4.50	6.00	13.50	3.50	1.00
	Quartet ↑	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Under this restricted-choice protocol, all seven backbones have zero *Quartet* on Reasoning. None answers every member of any of its 50 quartets correctly. At the same time, Reasoning *Accuracy* ranges from 23.00% to 38.50%. These item-level successes yield no complete-group success. The fixed-correctness analysis below relates this pattern to the distribution of item-level errors.

For Qwen3.5-9B, the 510 original-score successes include 294 correct replacement answers, 108 obsolete answers, and 108 other errors. Appendix C.3 decomposes these outcomes for all backbones and the adapter. Overall *Retention* is close to a uniform-distractor error reference; conditional outcomes reveal concentrations that aggregate rates can obscure, as detailed in Appendices B.1 and C.3.

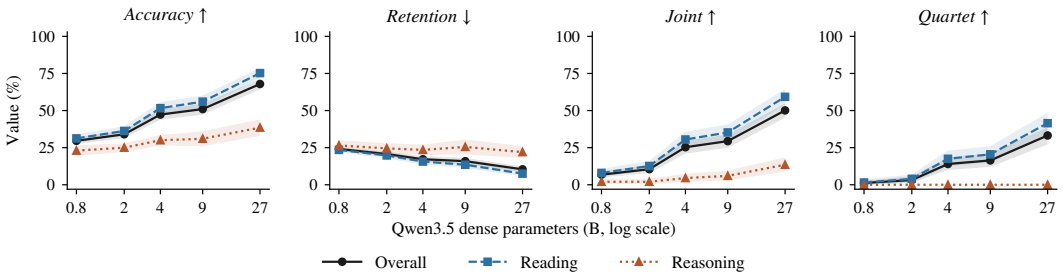

Figure 2: Qwen3.5 scaling under five-shot evaluation. Parameter count uses a logarithmic axis; shading shows 95% quartet-bootstrap intervals.

The seven backbones have the same Overall ranking by *Accuracy*, *Joint*, and *Quartet*. A fixed-correctness regrouping reference places actual *Quartet* between 2.05 percentage points below and 2.10 points above its mean across systems. Five of the seven backbones fall below their reference means. Gemma-2-9B falls below its central 95% reference range, while the adapter reaches the upper endpoint of its own range. Appendix B reports the reference design, full results, and uncertainty estimates.

Qwen3.5-9B with budget-limited thinking reaches Overall *Accuracy* of 54.0%, *Retention* of 15.4%, *Joint* of 35.0%, and *Quartet* of 23.2%. Reasoning *Quartet* remains zero. This supplementary setting uses chat-formatted inputs and sampled thinking before restricted-choice scoring. It evaluates these settings jointly, alongside the plain-text baseline. Table C1 reports component results and the accompanying protocol explains budget-limited answer extraction.

The five-shot API evaluation in Table 5 uses separate requests for each original and replacement item and scores the returned answer letters. It retains the same questions, options, and demonstrations, but omits the local prompt's instruction and reading-convention preamble. Reasoning settings also differ from the local baseline. We therefore report results separately for the two protocols. *Accuracy* ranges from 92.8% to 99.9% and *Joint* from 86.7% to 99.9%. Reasoning *Quartet* ranges from 72.0% to 100.0%, unlike the zero values under the open-weight restricted-choice setting. Near-complete success is consistent with the intended ability standard of correct reading and analysis of the supplied notation. Appendix C.2 gives component results and protocol details.

Table 5: Five-shot API results on the same test items, in percentages.

Model	***Accuracy*** ↑	***Retention*** ↓	***Joint*** ↑	***Quartet*** ↑
GPT-6 Astra	99.9	0.0	99.9	99.6
Fable 5	99.2	0.1	98.3	96.8
Opus 5	98.5	0.0	97.3	94.0
Sonnet 5	97.1	0.7	94.5	89.2
GPT-5.6 Sol	92.8	2.0	86.7	76.8

4.3 Robustness and External Diagnostics

Answer-preserving instruction rewording and choice permutation produce small changes under the two tested transformations. Qwen3.5-9B backbone *Accuracy* ranges from 51.00% to 52.10%; the adapter reaches 91.50% and 91.70%, with *Quartet* of 76.40% and 77.60%. Appendix C.4 gives the transformations and complete results.

A separate zero-shot diagnostic extends score replacement to verified external subsets. On 383 eligible SSMR-Bench items, *Accuracy* is 51.7%, *Retention* is 35.0%, and *Joint* is 11.2%, again distinguishing isolated correctness from correct updates. This diagnostic uses derived intervention subsets and reports each dataset separately. MusICA's original accuracy is near its majority-label baseline, which provides context for its replacement results. Appendix C.5 reports inclusion rules, results, and evaluation details.

4.4 Training Utility

Ordinary LoRA supervision tests whether the training split can improve score-dependent behavior (Hu et al., 2022). Cross-entropy covers only the answer label and end-of-sequence token; prompt tokens are masked. Quartet structure enters through the training examples. Both adapters use the ScoreQA training split for one epoch with a fixed configuration and final-checkpoint reporting; validation monitors training.

Table 6 compares the Qwen3.5-9B backbone and final ScoreQA adapter using matched items, demonstrations, and scoring. After adaptation, Reading *Accuracy* rises from 56.00% to 95.50% and *Quartet* from 20.50% to 84.00%. Reasoning rises from 31.00% and 0.00% to 74.00% and 44.00%, respectively. Overall *Joint* rises from 29.40% to 84.90%, while *Retention* falls from 15.90% to 3.00%. Gains extend beyond original-score answers to paired and complete-group correctness, accompanied by lower obsolete-answer retention. Under matched evaluation, MMLU changes from 80.20% to 80.24%, indicating broadly unchanged aggregate performance on this general-knowledge test.

Table 6: Qwen3.5-9B backbone / ScoreQA-adapted results, in percentages, including MMLU. Arrows indicate preferred directions; teal marks adapted values.

Scope	*Accuracy* ↑	*Retention* ↓	*Joint* ↑	*Quartet* ↑
Overall	51.00 / 91.20	15.90 / 3.00	29.40 / 84.90	16.40 / 76.00
Reading	56.00 / 95.50	13.50 / 1.62	35.25 / 91.38	20.50 / 84.00
Reasoning	31.00 / 74.00	25.50 / 8.50	6.00 / 59.00	0.00 / 44.00
MMLU	80.20 / 80.24		Not applicable	

Gains are not uniform across operations. Qwen's adapted event-type *Quartet* is 40.00%, compared with 100.00% for key signature, time signature, and voice clef. Llama-3.1-8B also improves, reaching 30.00% Reasoning *Quartet*, while event-type *Quartet* remains 4.00%. Its MMLU accuracy changes from 68.52% to 68.40%, a decrease of 0.12 percentage points under the same evaluation protocol. Appendix D preserves both models' full unit-level results. These single-run gains demonstrate learnability on source-disjoint test scores within the same task distribution.

4.5 Case Study

Table 7 follows the Reading and Reasoning examples in Table A2 through their assigned score replacements. It adds predictions from the same five-shot Qwen3.5-9B backbone and final adapter used in Table 6. Each pair keeps the question, choices, and demonstrations fixed. The evidence excerpts connect the reference answers to the notation in each window.

Table 7: Qwen3.5-9B predictions for two existing test examples. Questions are condensed and notation excerpted; models receive full windows.

Score	Key evidence	Reference	Backbone	ScoreQA-adapted
Reading				
Question. What is the written pitch at offset 0 in measure 1 of voice `V:1`?				
Choices. A. G3 B. G5 C. G#4 D. G4				
Original	`K:C`; treble; `^G`	C	A	C
Replacement	`K:G`; treble; `G`	D	D	D
Reasoning				
Question. What is the interval between the first two printed notes in measure 2 of voice `V:1`, ignoring grace notes?				
Choices. A. minor sixth B. minor third C. minor second D. minor seventh				
Original	`K:C`; treble; `E/ c/`	A	C	C
Replacement	`K:D`; treble; `C E2`	B	B	B

The Reading backbone changes from A to D and answers the replacement correctly, but misses the original; the adapted model answers C and D correctly. Accordingly, only the adapted model satisfies *Joint* for this Reading pair. For Reasoning, the replacement's `K:D` makes `C E2` denote C#4 followed by E4, a minor third rather than the original minor sixth. Both systems predict C, a minor second, on the original and B, the correct minor third, on the replacement. Both thus miss the original answer in this case. The example was selected as the first interval item in the frozen test set's identifier order, without conditioning on model success. These cases illustrate the distinction between changing answers and being correct under both scores.

5 Discussion

ScoreQA makes foundational score-reading and analysis abilities operational through explicit tasks, musical conventions, and correctness criteria. The MusicXML oracle and representation checks establish the reference answers; shared-question quartets then require models to distinguish evidence supporting different choices. ARJQ connects item correctness to success under paired scores and across complete quartets. This combination makes it possible to examine both whether a response is correct and how correctness is distributed across controlled score evidence.

The results support an attainment-oriented view of benchmark evaluation. Near-complete success in the five-shot API setting demonstrates that the covered tasks can be performed consistently. The open-weight restricted-choice results and adaptation study reveal variation across notation operations, including stronger state-reading performance and remaining errors in event type and musical comparisons. Each result is interpreted under its stated prompting and inference conditions. Together, the benchmark and its analyses provide a structured account of which tested operations models perform correctly and how the training split supports improvement.

6 Conclusion

ScoreQA evaluates defined score-reading and musical-analysis abilities using verifiable answers and score-grounded quartets. ARJQ summarizes correctness at item, pair, and complete-group levels alongside obsolete-answer retention. Open-weight restricted-choice baselines show incomplete success, while five-shot API results approach or reach the tested standard on some scopes. Together, the evaluations characterize attainment under their respective protocols. Complete correctness is the intended outcome of mastering the covered tasks.

Reproducibility Statement

Section 3 describes source separation, answer verification, quartet construction, and ARJQ. Appendices A through D provide benchmark statistics, construction examples, uncertainty analysis, evaluation settings, human-review results, and detailed adaptation results. The supplementary archive provides 1,112 items in 278 complete quartets, comprising 10% of the formal training, validation, and test splits. It includes the score windows, questions, choices, reference answers, replacement mappings, and data documentation. The complete dataset will be released after publication.

AI Use Statement

In this work, we did not use generative AI tools in paper writing.

References

Associated Board of the Royal Schools of Music. The ABRSM music theory syllabus from 2020: Grades 1 to 5, 2020. URL https://www.abrsm.org/sites/default/files/2024-01/music-theory-syllabus-outline-grades-1-5-from-2020.pdf.

Associated Board of the Royal Schools of Music. About music theory exams. https://www.abrsm.org/en-gb/music-theory/about-music-theory, n.d. Official qualification, syllabus, and practice-resource overview. Accessed 15 September 2026.

Congren Dai, Yue Yang, Krinos Li, Huichi Zhou, Shijie Liang, Bo Zhang, Enyang Liu, Ge Jin, Hongran An, Haosen Zhang, Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Xiaobing Li, and Maosong Sun. Musical score understanding benchmark: Evaluating large language models' comprehension of complete musical scores. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10777–10799. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.493. URL https://aclanthology.org/2026.acl-long.493/.

Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, et al. ComposerX: Multi-agent symbolic music composition with LLMs. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pp. 669–679, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Phillip Long, Zachary Novack, Taylor Berg-Kirkpatrick, and Julian McAuley. PDMX: A large-scale public domain MusicXML dataset for symbolic music processing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890217.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, et al. MuPT: A generative symbolic music pretrained transformer. In *International Conference on Learning Representations*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Tomáš Sourada, Katia Vendrame, and Jan Hajič, Jr. Music I care about: Automated multimodal benchmarking of LLM music perception skills on (almost) any music. *arXiv preprint arXiv:2607.06015*, 2026.

Bob L. Sturm, João Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. Music transcription modelling and composition using deep learning. *arXiv preprint arXiv:1604.08723*, 2016.

Chris Walshaw. The ABC music standard 2.1. `https://abcnotation.com/wiki/abc:standard:v2.1`, 2011.

Zhilin Wang, Zhe Yang, Yun Luo, Yafu Li, Xiaoye Qu, Ziqian Qiao, Haoran Zhang, Runzhe Zhan, Derek F. Wong, Jizhe Zhou, and Yu Cheng. Towards an AI musician: Synthesizing sheet music problems for musical reasoning. *arXiv preprint arXiv:2509.04059*, 2025.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*, 2016.

Shangda Wu, Yashan Wang, Xiaobing Li, Feng Yu, and Maosong Sun. MelodyT5: A unified score-to-score transformer for symbolic music processing. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, 2024.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Jiahao Zhao, Yunjia Li, Wei Li, and Kazuyoshi Yoshii. ABC-Eval: Benchmarking large language models on symbolic music understanding and instruction following. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026. doi: 10.1109/ICASSP55912.2026.11463289.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.

A Benchmark Composition, Construction, and Human Review

A.1 Source Coverage and Sampling Structure

Table A1 reports the retained source audit. Source groups collect related files before splitting; distinct ABC windows are counted by exact content hashes. Multiple items can share a window, and one source file can supply multiple windows. Counts therefore describe different units rather than independent samples. All retained items use sources marked public domain or CC0 in the source metadata. File, source-group, and ABC-content intersections are empty across splits.

Table A1: Source and representation coverage by split. Counts are distinct within each split.

Split	Source files	Source groups	ABC windows
Train	3,842	3,790	3,842
Validation	57	53	91
Test	442	433	442
Demonstrations	19	18	19

The benchmark allocation is deliberately balanced by task rather than sampled to reproduce the frequency of musical concepts in a repertoire. Each Reading unit contributes the same number of items; pooled Reasoning contributes twice that number. Consequently, Overall ARJQ combines Reading and Reasoning with weights of 80% and 20%, respectively, at both item and quartet levels. Component-level results describe performance within this fixed mixture.

Every complete quartet contains one A, B, C, and D reference label. The training, validation, and test splits therefore contain 2,500, 30, and 250 items per label, respectively. This balance follows directly from quartet construction. The 45 demonstrations are independent items. Source files, windows, items, and quartets define distinct sampling units. The 1,000 test items reuse 442 ABC windows, and related files belong to 433 source groups. Within-split reuse can induce dependence across questions even though cross-split source-group and content intersections are empty.

Figure 3 characterizes the retained score windows, counting each distinct ABC text once per split rather than weighting it by the number of questions. The test medians are four measures, two active voices, 37 events, and 215 ABC characters. The corresponding interquartile ranges are 3 to 4, 2 to 4, 20 to 61.75, and 136.25 to 330. An active voice contains at least one visible nongrace note or rest. Events include rests and count a chord as one event and a tied continuation separately; grace notes and invisible rests are excluded. Character counts include ABC headers and newlines. These statistics characterize the displayed notation.

Test windows cover 14 key signatures, 17 time signatures or none, and 10 clef types. Clef coverage counts presence anywhere in any voice, so one window can contribute to several bars. The figure pools tasks within each split; Reasoning remains a single category in all task-level summaries. The demonstration distribution summarizes its 19 distinct windows.

A.2 Task Grounding and Worked Construction

The ABRSM Grades 1 to 5 syllabus grounds the task inventory in specific notation and theory topics (Associated Board of the Royal Schools of Music, 2020). Grade 1 introduces note and rest values, ties, meter, clefs, accidentals, keys, and tonic triads, corresponding to foundational Reading operations. Grades 3 to 5 extend clefs, keys, and rhythmic notation. Grade 4 includes intervals within an octave, and Grade 5 includes concert-pitch transposition, motivating comparisons and transformations in Reasoning. ScoreQA organizes these notation and theory topics into its own window-based tasks, cross-voice queries, and answer-balanced quartets.

Table A2 follows the construction from target selection through verification. The Reading quartet asks for written pitch at the same position, with four windows supporting G3, G5, G#4, and G4. The Reasoning example instead requires a relation between two identified notes. Under `K:C`, the first two printed notes of measure 2, `E` and `c`, denote E4 and C5 and form a minor sixth. It therefore illustrates the additional comparison rather than pitch identification alone. Section 4.5 examines model responses to these examples and their assigned replacements.

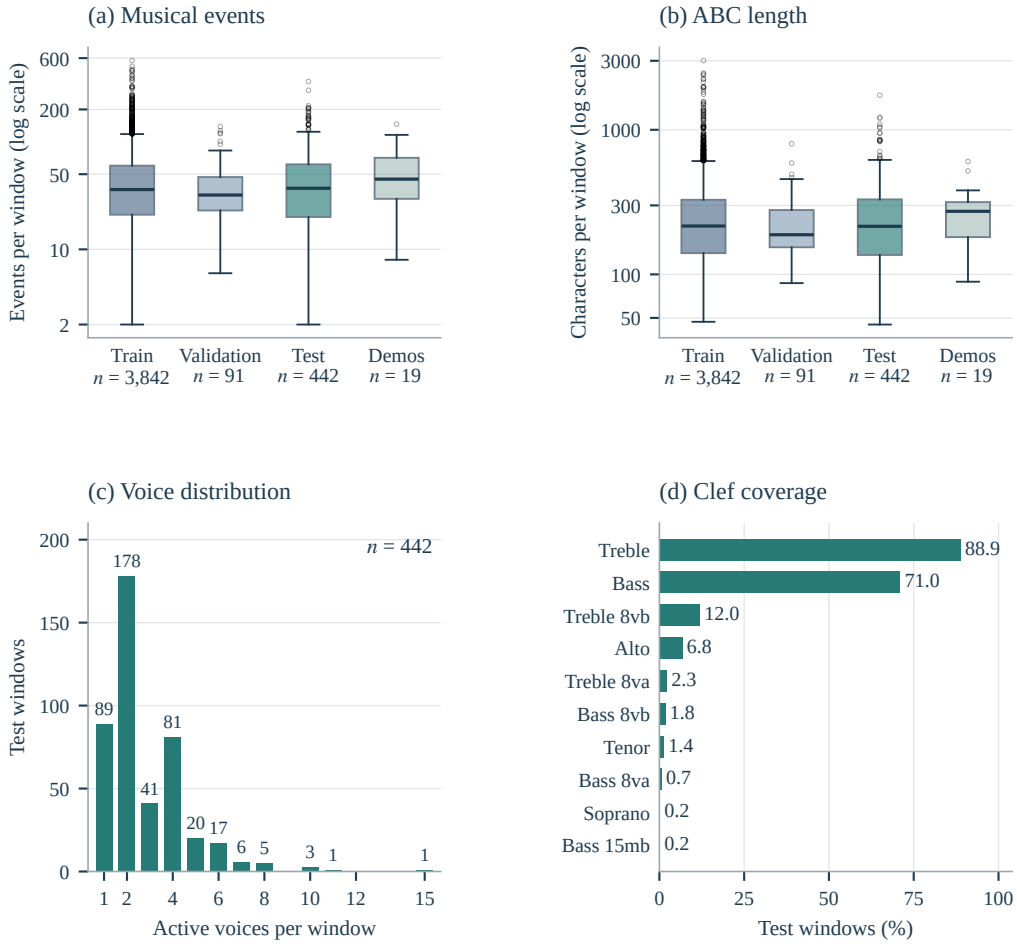

Figure 3: Score-window distributions by split and test-set notation coverage. Boxes show medians and interquartile ranges, with 1.5-IQR whiskers and all outliers on logarithmic axes. Clef percentages use 442 test windows and may sum to more than 100%.

A.3 Independent Human Review

Two human reviewers with musical backgrounds independently answered the same 96 test questions without LLM assistance. The sample comprised 24 complete quartets drawn with seed 20260921 under task-coverage and source-separation constraints. Twelve quartets covered all eight Reading units, with one or two quartets per unit; twelve covered Reasoning, with two quartets per underlying operation family. The 96 items came from distinct source groups. Items were shuffled and assigned anonymous question numbers. The review materials retained the original ABC scores, questions, and choice order, without reference answers, oracle outputs, model predictions, or quartet identifiers. Reviewers supplied one answer letter per question.

Table A3 separates agreement with the reference answers from agreement between reviewers. The reviewers differed on six questions, with Reviewer 2 matching the reference on all six. Both selected the same non-reference option on one additional question involving a key-signature accidental. The original responses and reference answers were retained unchanged. Human answer agreement complements the programmatic verification reported in Section 3.1.

The review sample gives Reading and Reasoning equal weight, unlike the 80/20 mixture in the full test set. The reported human agreement rates are therefore specific to this coverage-oriented sample.

Table A2: Test-set Reading construction and Reasoning examples. Only target notation is shown; models receive complete windows. Reading members come from distinct source groups.

Reading construction

1. Locate the target and instantiate the question

Question: In ABC voice `V:1`, measure 1, what is the written pitch of the note at offset 0 quarter notes from the start of the measure?

2. Compute answers and match four source windows

Member	Key	Voice and target note	Oracle	Final label
A	`K:G V:1 treble`; first note	`G,4`	G3	A
B	`K:F V:1 treble`; first note	`vg`	G5	B
C	`K:C V:1 treble`; first note	`^G`	G#4	C
D	`K:G V:1 treble`; first note	`G`	G4	D

3. Form the shared choices and assign labels

Choices: A. G3 B. G5 C. G#4 D. G4

4. Verify the displayed evidence and complete the quartet

Each displayed ABC answer must match the MusicXML oracle and exactly one choice. The four members share the question and choices, with each label correct once.

Reasoning: comparing two pitches

Question: In ABC voice `V:1`, measure 2, take note 1 and note 2 of the bar, counting printed notes, ignoring grace notes; a chord that starts together counts as one note and a note tied from an earlier note still counts. What is the interval between them?

Choices: A. minor sixth B. minor third C. minor second D. minor seventh

Member	Key	Voice and target notes	Oracle	Label
A	`K:C V:1 treble`; bar 2,	`E/ c/`	minor sixth	A

Table A3: Independent human review of the same 96 test items.

Comparison	Matching answers	Agreement (%)
Reviewer 1 vs. reference	89/96	92.71
Reviewer 2 vs. reference	95/96	98.96
Reviewer 1 vs. Reviewer 2	90/96	93.75

B Statistical Properties of the Evaluation

B.1 Label Balance and Metric Resolution

For a deterministic score-blind predictor, all four members of a quartet receive the same answer. Exactly one original member and one replacement member have that label as their reference, but they are different members because replacement advances the label. Hence the construction implies

$$\textit{Accuracy} = \textit{Retention} = \frac{1}{4}, \qquad \textit{Joint} = \textit{Quartet} = 0. \tag{4}$$

This gives the analytical no-score reference. Under a different hypothetical predictor that guesses independently and uniformly on distinct score prompts, the expectations are

$$\mathbb{E}[\textit{Accuracy}] = \mathbb{E}[\textit{Retention}] = \frac{1}{4}, \qquad \mathbb{E}[\textit{Joint}] = \frac{1}{16}, \qquad \mathbb{E}[\textit{Quartet}] = \frac{1}{256}. \tag{5}$$

The two analytical references distinguish deterministic responses to identical score-removed prompts from independent guesses on distinct score prompts.

An error-conditioned reference instead holds replacement correctness fixed and assumes that wrong responses are uniform over the three distractors. Exactly one distractor is the original correct label. If p_{swap} denotes replacement-answer accuracy as a proportion, expected *Retention* under this assumption is $(1 - p_{\text{swap}})/3$. For deterministic donor scoring, this reduces to $(1 - \textit{Accuracy})/3$. For independently sampled API responses, the reference uses replacement accuracy, not original-score

Accuracy. This reference describes error allocation under the uniform-distractor assumption. On the originally correct subset, the analogous reference count is one third of the replacement errors in that subset, provided the same uniform-error assumption holds conditionally.

Observed metric resolution follows the declared denominators. Overall item metrics change in increments of 0.1 percentage points, while one successful quartet changes Overall *Quartet* by 0.4 points. Reading, Reasoning, and individual Reading units contain 200, 50, and 25 test quartets, respectively, so their *Quartet* increments are 0.5, 2, and 4 points. Fine numerical differences on small categories should therefore be read alongside the underlying group counts.

B.2 Uncertainty Estimates

The 95% intervals in Figure 2 use 10,000 quartet-cluster bootstrap draws with seed 21. Each draw samples whole quartets with replacement from the relevant pool, retaining the original number of quartets and all four members of each selected group. Overall, Reading, and Reasoning intervals use their respective pools of 250, 200, and 50 quartets. Overall resampling samples from the combined pool, allowing its Reading/Reasoning mixture to vary. The 2.5th and 97.5th percentile endpoints summarize variation under quartet resampling, conditional on the evaluated model, reference answers, and protocol.

Table B1 provides the existing Overall intervals for the seven backbones and the Qwen adapter. Each interval summarizes one system's empirical variation under quartet resampling. At a boundary such as zero successful Reasoning quartets, every resample also has zero successes, yielding a zero-width empirical interval. Resampling whole quartets preserves within-quartet dependence. Across task units, shared source groups and windows connect the 250 test quartets into one component; within an individual task unit, distinct quartets share no source groups. Templates are also reused within tasks. These source and template links remain outside the resampling unit, so the intervals are conditional quartet-bootstrap summaries rather than source-population uncertainty estimates.

Table B1: Marginal 95% quartet-bootstrap intervals for each system's Overall ARJQ, in percent.

Model	*Accuracy*	*Retention*	*Joint*	*Quartet*
Qwen3.5-0.8B	[27.5, 31.8]	[22.6, 25.6]	[4.9, 9.0]	[0.0, 2.8]
Qwen3.5-2B	[31.4, 36.7]	[19.0, 22.4]	[7.9, 13.4]	[1.2, 5.6]
Qwen3.5-4B	[43.8, 50.9]	[15.3, 19.2]	[21.0, 29.7]	[10.0, 18.4]
Qwen3.5-9B	[47.3, 54.6]	[13.8, 18.1]	[24.9, 33.8]	[12.0, 21.2]
Qwen3.5-27B	[64.3, 71.3]	[8.6, 12.3]	[45.2, 54.8]	[27.2, 38.8]
Llama-3.1-8B	[38.7, 45.2]	[20.1, 23.8]	[14.9, 22.8]	[6.8, 14.4]
Gemma-2-9B	[37.5, 43.8]	[17.7, 21.4]	[13.7, 21.3]	[5.6, 12.4]
Qwen3.5-9B (ScoreQA-adapted)	[88.8, 93.4]	[2.0, 4.1]	[81.2, 88.4]	[70.4, 81.2]

B.3 Fixed-Correctness Regrouping

Quartet tests whether correct answers cover all four scores sharing a question. To relate it to item accuracy, each item's original correctness is held fixed, and quartets are randomly regrouped within each of the eight Reading units and the pooled Reasoning category. Each new group contains one item per answer label. The procedure is repeated 10,000 times while preserving category and label counts. Actual *Quartet* differs from the regrouping mean by between 2.05 percentage points lower and 2.10 points higher across systems. This reference describes how correct answers are distributed across original and random groups, with Reasoning retained as one pooled category. *Quartet* reports strict group success, while the reference contextualizes that success given the observed item-correctness pattern.

Table B2 reports the full reference summaries. The central reference range spans the 2.5th to 97.5th percentiles over regroupings, conditional on observed item correctness and one item per label within each task category. This reference varies group membership, whereas the cluster bootstrap above resamples the original groups.

Table B2: Actual *Quartet* and its fixed-correctness regrouping reference. Values are percentages; differences are percentage points.

Model	Actual	Mean	Central 95%	Difference
Qwen3.5-0.8B	1.20	1.11	[0.40, 2.00]	0.09
Qwen3.5-2B	3.20	2.32	[1.20, 3.60]	0.88
Qwen3.5-4B	14.00	15.34	[14.00, 16.80]	-1.34
Qwen3.5-9B	16.40	17.51	[15.60, 19.60]	-1.11
Qwen3.5-27B	33.20	35.25	[33.20, 37.60]	-2.05
Llama-3.1-8B	10.40	11.85	[10.40, 13.60]	-1.45
Gemma-2-9B	8.80	10.78	[9.60, 12.40]	-1.98
Qwen3.5-9B (ScoreQA-adapted)	76.00	73.90	[72.00, 76.00]	2.10

C Evaluation Settings and Supplementary Diagnostics

C.1 Prompts and Scoring

The open-weight evaluation uses five fixed demonstrations for each of the eight Reading units and the pooled Reasoning category, drawn from the separate demonstration split. The prompt starts with a multiple-choice instruction and the reading conventions in Section 3.1. Each demonstration then supplies an ABC window, question, choices, and answer; the target uses the same layout but ends at `Answer:`. Within a quartet, all text outside the target score is identical. No chat template or extra special tokens are added. The scorer compares the final-position logits of the single-token strings `A`, `B`, `C`, and `D`, selecting the largest value with ties resolved in that order. Inputs use left padding, bfloat16 inference, and scaled dot-product attention. The evaluation length guard is 16,384 tokens and rejects overlength inputs rather than truncating or skipping them. Every reported backbone run completes all 1,000 items. The 4,096-token training block is a separate setting.

For Qwen3.5-9B with budget-limited thinking, we retain the same five-shot content, reading conventions, questions, choices, and score replacements. The target's final answer cue is removed before applying the model's chat template with thinking enabled. Thinking is sampled independently for each item and condition under a fixed generation budget. At the budget limit, thinking is forcibly closed if it has not ended naturally. We then append `Answer:` and apply the same raw next-token comparison over the four answer letters, without sampling the final answer. This setting combines chat formatting, sampled thinking with budget-limited closure, and restricted-choice answer extraction.

Table C1 reports one sampled run covering all 1,000 original items and their replacements, with no missing responses. Original and replacement calls use separately seeded thinking, so replacement accuracy can differ from original *Accuracy*. Improvements over the plain-text baseline are concentrated in key signatures and voice clefs, while event duration, event type, and simultaneous-pitch accuracy decline.

Table C1: Qwen3.5-9B with budget-limited thinking, in percentages. The unadapted model uses restricted answer-letter scoring after thinking.

Scope	*Accuracy* ↑	*Retention* ↓	*Joint* ↑	*Quartet* ↑
Overall	54.00	15.40	35.00	23.20
Reading	59.25	13.75	41.88	29.00
Reasoning	33.00	22.00	7.50	0.00

C.2 API Evaluation and Component Results

Five models are evaluated through independent five-shot API requests on all 1,000 original-score items and their 1,000 assigned replacements. Table 5 uses consistently formatted display names for the supplied models. Each request contains five answered demonstrations followed by one unan-

swered target question. The questions, option order, demonstrations, and quartet mapping are unchanged from the frozen test. Unlike the local prompt, the exported request has no general task instruction or reading-convention preamble. A replacement request does not receive the original response or its reasoning. Returned letters are aligned by item identifier and scored against the original and donor labels using the same ARJQ definitions. The frozen test questions are submitted to external services. Reasoning settings and token budgets vary across the API and local evaluations, whose results are reported under their respective inference protocols.

Each answer sheet contains 1,000 unique item identifiers without missing answers. Table C2 reports Reading and pooled Reasoning separately. Correctness against the replacement score's new label is 100.0%, 99.1%, 98.8%, 97.3%, and 93.1% in the model order shown. Every replacement input is also the original input of its donor member. Separate API calls therefore repeat these inputs under the replacement assignment. Sampling permits differences between the observed original and replacement accuracies. Low *Retention* together with high *Joint* describes correctness across both score conditions.

Table C2: API Reading and Reasoning results, in percentages. Reading contains 800 items and 200 quartets; Reasoning contains 200 items and 50 quartets.

Scope	Model	*Accuracy* ↑	*Retention* ↓	*Joint* ↑	*Quartet* ↑
Reading	GPT-6 Astra	99.88	0.00	99.88	99.50
	Fable 5	99.62	0.00	99.25	98.50
	Opus 5	98.75	0.00	97.75	95.00
	Sonnet 5	97.75	0.25	96.50	91.50
	GPT-5.6 Sol	93.12	1.38	89.12	78.00
Reasoning	GPT-6 Astra	100.00	0.00	100.00	100.00
	Fable 5	97.50	0.50	94.50	90.00
	Opus 5	97.50	0.00	95.50	90.00
	Sonnet 5	94.50	2.50	86.50	80.00
	GPT-5.6 Sol	91.50	4.50	77.00	72.00

C.3 Replacement Outcomes for Originally Correct Items

Table C3 decomposes existing five-shot predictions to distinguish behavior after an originally correct answer. For each model, only items answered correctly with the original score are considered. Replacement responses are divided into the correct new answer, the obsolete original answer, and another wrong answer. Counts sum to the model's original-correct total. Dividing the new-correct count by the full test size gives *Joint*. Old-answer counts here concern only the originally correct subset, while full-test *Retention* uses all items. This analysis uses existing predictions to describe the composition of replacement outcomes.

Table C3: Replacement outcomes within each model's originally correct subset.

Model	Original correct	New answer correct	Old answer retained	Other error
Qwen3.5-0.8B	296	68	194	34
Qwen3.5-2B	340	105	168	67
Qwen3.5-4B	473	253	133	87
Qwen3.5-9B	510	294	108	108
Qwen3.5-27B	679	501	74	104
Llama-3.1-8B	419	188	174	57
Gemma-2-9B	406	174	155	77
Qwen3.5-9B (ScoreQA-adapted)	912	849	26	37

Qwen3.5-0.8B answers 296 original items correctly but retains the old answer on 194 after replacement, or 65.54% of its originally correct subset. These are original-score successes paired with incorrect replacement answers. Among its 228 replacement errors in this subset, uniform allocation across the three distractors would yield 76 old-answer responses, rather than the observed 194. This conditional concentration is not apparent from the aggregate *Retention* alone. Of the Qwen3.5-9B backbone's 510 originally correct items, 294 receive the correct new answer, 108 retain the old answer, and 108 change to another error. For the adapter's 912 originally correct items, the corresponding counts are 849, 26, and 37. These conditional subsets differ, so the paired adaptation comparison remains the full-test comparison in Table 6. The decomposition distinguishes changed-but-wrong answers from correct new answers and obsolete answers; the analytical uniform-error reference in Appendix B.1 contextualizes their allocation.

C.4 Answer-Preserving Prompt Controls

Prompt analysis complements score replacement by testing sensitivity to wording and choice positions when musical answers remain unchanged. Instruction rewording preserves the question, score, and choices. Choice permutation consistently remaps labels in both target items and demonstrations. Each transformation is shared across all members of a quartet, preserving balance. In Table C4, the adapter reaches 91.50% and 91.70% *Accuracy* under instruction rewording and choice permutation, respectively, with *Quartet* of 76.40% and 77.60%. Changes from the original prompt do not exceed 0.50 and 1.60 percentage points for these two metrics. Backbone *Accuracy* ranges from 51.00% to 52.10% across conditions. The observed changes are small under these two fixed transformations.

Table C4: Overall ARJQ for Qwen3.5-9B before and after ScoreQA adaptation under answer-preserving prompt changes. Values are percentages.

System	Prompt	*Accuracy* ↑	*Retention* ↓	*Joint* ↑	*Quartet* ↑
Backbone	Original	51.00	15.90	29.40	16.40
	Reworded instructions	51.00	16.40	30.00	15.60
	Permuted choices	52.10	17.50	29.30	15.60
ScoreQA-adapted	Original	91.20	3.00	84.90	76.00
	Reworded instructions	91.50	2.60	85.40	76.40
	Permuted choices	91.70	2.80	85.90	77.60

C.5 Derived External Subsets

Cross-benchmark analysis tests whether replacement diagnostics also apply to derived subsets of other resources. All conditions use Qwen3.5-9B before ScoreQA adaptation, with zero-shot plain-text prompts and restricted answer-letter scoring. Zero-shot evaluation avoids mixing demonstration sources across datasets; it differs from the five-shot protocol in Tables 4 and 6. An external item is eligible only when a deterministic verifier reproduces its published answer and a valid notation change makes another existing choice correct. Questions and choice sets stay fixed. Replacements and displayed choice order are selected before inference without model predictions. The SSMR-Bench subset changes key signatures or notes in chord-root and interval questions; 383 of 399 candidates meet the inclusion rules. The MusICA subset follows its note-indexing rules and five-choice construction, including *none of the above*, using replacement scores from different sources. Checks on displayed scores exclude rendering inconsistencies, and duplicate items are removed before final evaluation. The derived subsets are evaluated with this common intervention protocol.

Table C5 illustrates why original-answer correctness and response to changed evidence should be reported separately. The SSMR-Bench subset has 51.7% *Accuracy*, but 35.0% of predictions after replacement still match the obsolete original answer, and only 11.2% of items are correct under both scores. These measurements show that the diagnostic can inform eligible external tasks. MusICA's 23.1% *Accuracy* is close to its 21.1% majority-label baseline, which constrains interpretation of its replacement responses as evidence of score reliance. External subsets do not form complete answer-balanced quartets, so *Quartet* is not reported. ScoreQA reaches 51.5% *Accuracy* and 29.4% *Joint* in its separate zero-shot run. Results are reported descriptively within each derived subset.

Table C5: Zero-shot replacement diagnostics with the Qwen3.5-9B backbone on fixed derived subsets, in percentages.

Dataset	Items	*Accuracy* ↑	*Retention* ↓	*Joint* ↑
ScoreQA	1,000	51.5	15.1	29.4
SSMR-Bench	383	51.7	35.0	11.2
MusICA-MetaBench	655	23.1	20.9	2.3

D Training Configuration and Detailed Adaptation Results

Both adaptation runs use seed 21, LoRA rank 64, alpha 128, and dropout 0.05. Each processes 10,000 items for one epoch with 625 optimizer updates. The effective batch size is 16, obtained from one item per GPU, four GPUs, and four accumulation steps. AdamW uses zero weight decay, gradient clipping at 1, 50 warmup steps, and cosine decay from a peak learning rate of 5×10^{-5}. Training uses bfloat16 and gradient checkpointing. LoRA covers attention and feed-forward projections, including Qwen's linear-attention projections. Training targets contain a leading-space answer label and EOS. Training prompts sample zero to five demonstrations per item, so exact label balance applies to the shared question and choices, not every complete training prompt. Evaluation fixes the demonstration prefix within each quartet. Full-score and replacement validation runs every 125 steps for monitoring; all results use the final one-epoch checkpoint.

Table D1 expands the Qwen summary in Table 6, retaining every Reading unit to characterize task-dependent behavior.

Reading-unit results clarify where adaptation helps. Before adaptation, the time-signature unit has 100.00% *Accuracy*, *Joint*, and *Quartet*, while written pitch and simultaneous pitches have zero *Quartet*. After adaptation, the latter two reach 68.00% and 80.00% *Quartet*. Key signature, time signature, and voice clef each reach 100.00%. Event type remains weaker, with 81.00% *Accuracy* and 40.00% *Quartet*. Unit and group metrics show how gains vary across notation operations.

Table D1: Qwen3.5-9B backbone / ScoreQA-adapted results, in percentages, including MMLU. Arrows indicate preferred directions; teal marks adapted values.

Scope	*Accuracy* ↑	*Retention* ↓	*Joint* ↑	*Quartet* ↑
Overall	51.00 / 91.20	15.90 / 3.00	29.40 / 84.90	16.40 / 76.00
Reading	56.00 / 95.50	13.50 / 1.62	35.25 / 91.38	20.50 / 84.00
Written pitch	33.00 / 92.00	21.00 / 5.00	9.00 / 84.00	0.00 / 68.00
Event duration	66.00 / 99.00	12.00 / 0.00	44.00 / 98.00	12.00 / 96.00
Event type	34.00 / 81.00	16.00 / 5.00	10.00 / 65.00	0.00 / 40.00
Key signature	46.00 / 100.00	15.00 / 0.00	19.00 / 100.00	8.00 / 100.00
Time signature	100.00 / 100.00	0.00 / 0.00	100.00 / 100.00	100.00 / 100.00
Chord pitches	52.00 / 97.00	18.00 / 2.00	26.00 / 94.00	4.00 / 88.00
Simultaneous pitches	35.00 / 95.00	20.00 / 1.00	7.00 / 90.00	0.00 / 80.00
Voice clef	82.00 / 100.00	6.00 / 0.00	67.00 / 100.00	40.00 / 100.00
Reasoning	31.00 / 74.00	25.50 / 8.50	6.00 / 59.00	0.00 / 44.00
MMLU	80.20 / 80.24		Not applicable	

Llama-3.1-8B uses the same training protocol, with LoRA on its attention and feed-forward projections. Table D2 shows gains in both components, extending the findings beyond Qwen. Reasoning *Quartet* rises from zero to 30.00%. Reading performance remains uneven. Event type identification reaches 53.00% *Accuracy* and 4.00% *Quartet*, while key signature, time signature, and voice clef reach 100.00% on both metrics in this test set. The time-signature backbone already reaches 97.00% *Accuracy*, leaving little room for improvement. On the same 14,042-item MMLU test, the backbone answers 9,622 items correctly and the adapter 9,605, giving accuracies of 68.52% and 68.40%. Both evaluations use subject-matched five-shot prompts and restricted answer-letter scoring without generated thinking.

Table D2: Llama-3.1-8B backbone / ScoreQA-adapted results, in percentages, including MMLU. Arrows indicate preferred directions; teal marks adapted values.

Scope	***Accuracy*** ↑	***Retention*** ↓	***Joint*** ↑	***Quartet*** ↑
Overall	41.90 / 82.20	21.90 / 5.10	18.80 / 71.70	10.40 / 59.60
Reading	45.25 / 87.38	20.88 / 3.88	22.62 / 78.50	13.00 / 67.00
Written pitch	28.00 / 85.00	27.00 / 7.00	1.00 / 74.00	0.00 / 56.00
Event duration	49.00 / 84.00	23.00 / 4.00	23.00 / 70.00	4.00 / 52.00
Event type	25.00 / 53.00	29.00 / 12.00	1.00 / 27.00	0.00 / 4.00
Key signature	27.00 / 100.00	25.00 / 0.00	1.00 / 100.00	0.00 / 100.00
Time signature	97.00 / 100.00	2.00 / 0.00	94.00 / 100.00	88.00 / 100.00
Chord pitches	35.00 / 94.00	22.00 / 3.00	9.00 / 88.00	0.00 / 76.00
Simultaneous pitches	32.00 / 83.00	23.00 / 5.00	6.00 / 69.00	0.00 / 48.00
Voice clef	69.00 / 100.00	16.00 / 0.00	46.00 / 100.00	12.00 / 100.00
Reasoning	28.50 / 61.50	26.00 / 10.00	3.50 / 44.50	0.00 / 30.00
MMLU	68.52 / 68.40		Not applicable	