WildScore: Benchmarking MLLMs in-the-Wild Symbolic Music Reasoning

Gagan Mundada[1*] **Yash Vishe**[1*] **Amit Namburi**[1] **Xin Xu**[1]
Zachary Novack[1] **Julian McAuley**[1] **Junda Wu**[1]
[1]University of California, San Diego
{gmundada,yvishe,anamburi,xinxucs,znovack,jmcauley,juw069}@ucsd.edu

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities across various vision-language tasks. However, their reasoning abilities in the *multimodal symbolic music* domain remain largely unexplored. We introduce **WildScore**, the first in-the-wild multimodal symbolic music reasoning and analysis benchmark, designed to evaluate MLLMs' capacity to interpret real-world music scores and answer complex musicological queries. Each instance in WildScore is sourced from genuine musical compositions and accompanied by authentic user-generated questions and discussions, capturing the intricacies of practical music analysis. To facilitate a comprehensive evaluation, we propose a systematic taxonomy, comprising both high-level and fine-grained musicological ontologies. Furthermore, we frame complex music reasoning as multiple-choice question answering, enabling controlled and scalable assessment of MLLMs' symbolic music understanding. Empirical benchmarking of state-of-the-art MLLMs on WildScore reveals intriguing patterns in their visual-symbolic reasoning, uncovering both promising directions and persistent challenges for MLLMs in symbolic music reasoning and analysis. We release the dataset[1] and code[2].

1 Introduction

Multimodal Large Language Models (MLLMs) have recently advanced on visual question answering (Yan et al., 2024; Liu et al., 2023a), document understanding (Luo et al., 2024; Zhu et al., 2024; Wu et al., 2025d), visual navigation (Wu et al., 2025a; Wang et al., 2025; Wu et al., 2024c), and recommendation (Wu et al., 2024b; Huang et al., 2025). Despite these advances, the real-world applicability of MLLMs in symbolic music analysis and reasoning remains underexplored. Symbolic music reasoning uniquely combines dense visual symbolism with rich, domain-specific semantics (Yuan et al., 2024), posing challenges that extend beyond conventional image-text benchmarks (Fu et al., 2024; Yu et al., 2023). While there has been some limited work in evaluating LLMs on symbolic music tasks (Yuan et al., 2024), such work has only considered unimodal LLMs, where the symbolic music has been converted to text, on pedagogical-style test questions, which calls into question such benchmarks' ability to evaluate diverse reasoning performance. On the other hand, existing symbolic music datasets, like MusicNet (Thickstun et al., 2017a) and MAESTRO (Hawthorne et al., 2019), focus on aligned transcription or generation based on specific model architectures, which makes them unaligned with reasoning tasks or interfacing with larger text-based models. Unlike prior benchmarks that focus on unimodal audio analysis, OMR, or symbolic transcription, there remains no standardized evaluation for complex reasoning and analysis over symbolic music based on multimodal context, where understanding often hinges on multi-step deduction, ambiguity resolution, and integration of notation, structure, and expressive intent (Czajka et al., 2024a).

In this work, we present **WildScore**, the first multimodal symbolic music reasoning benchmark constructed from in-the-wild data. WildScore comprises real music scores by actual composers, paired with user-generated questions and discussions sourced from public forums. Many real-world queries require integrating several musical reasoning steps, including identifying notational symbols, interpreting harmonic progressions, and contextualizing expressive markings, which demand the need for MLLMs that can perform compositional and context-aware multimodal reason-

[*]These authors contributed equally to this work.
[1]https://huggingface.co/datasets/GM77/WildScore
[2]https://github.com/GaganVM/WildScore

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16847–16863
November 4–9, Suzhou, China, 2025

ing. This collection reflects the authentic diversity and complexity of symbolic music interpretation as it occurs in real-world discourse, and demands nuanced reasoning about notation, structure, and musical intent (Xu et al., 2024; Surana et al., 2022).

To enable a comprehensive and interpretable evaluation, we introduce a systematic taxonomy that covers both broad and detailed facets of music theory, including Harmony & Tonality, Rhythm & Meter, Expression & Performance, Texture and Form. This systematic taxonomy guides dataset curation and provides fine-grained analysis of MLLMs' strengths and limitations across musicological concepts.

To overcome the inherent ambiguity and subjectivity in open-ended musicological questions, we further propose to formulate symbolic music reasoning as a multiple-choice question answering (QA) problem.

Each WildScore instance presents a score image, an LLM-generated MCQ based on a real community submission, and several plausible answer candidates derived from the post's annotated ground truth. We illustrate the overview of our dataset in Figure 1. This figure displays the different high-level categories and subcategories, highlighting the range of musical topics and question types included in WildScore.

This controlled QA formulation allows for rigorous benchmarking, scalable annotation, and automatic evaluation while maintaining the authenticity of real-world musicological challenges.

Our empirical benchmarking of state-of-the-art MLLMs on WildScore (see Section 4) reveals that even widely used and popular models exhibit inconsistent accuracy across various musical reasoning tasks. Although recent vision–language models have demonstrated strong performance on prominent multimodal benchmarks (Ishmam et al., 2025) (Chen and Wu, 2024), they often are premature when faced with the deep musical abstractions and context-sensitive inferences required by real-world score interpretation. These observations point to a substantial gap that future multimodal models must close in order to fully capture the complexity of symbolic music analysis.

We summarize our contributions as follows:

- We introduce WildScore, the first in-the-wild symbolic music reasoning benchmark, grounded in real music scores and authentic expert questions.
- We propose a systematic, multi-level taxonomy for musicological reasoning, supporting comprehensive evaluation of MLLMs.
- We formulate complex symbolic music reasoning as multiple-choice QA, enabling controlled and scalable benchmarking.
- We conduct extensive empirical studies, providing the first insights into MLLMs' symbolic music reasoning capabilities and highlighting challenges for future research.

2 Related Work

2.1 Symbolic Music Understanding and Benchmarks

Symbolic music understanding has traditionally been evaluated using clean, structured datasets such as MusicNet (Thickstun et al., 2017a), NES-MDB (Donahue et al., 2018), and MAESTRO (Hawthorne et al., 2019). These datasets align audio with symbolic formats to facilitate tasks like transcription and generation. However, they reflect highly curated environments, lacking the variability, ambiguity, and informal nature of user-generated content. Other symbolic corpora like MusicScore (Lin et al., 2024) or Lakh MIDI (Raffel, 2016) further extend coverage but remain either score-centric or MIDI-based without real-world contextual grounding. Recent efforts like MusicTheoryBench (Czajka et al., 2024b) introduce theory-centric evaluations, but they rely on expert-curated questions in controlled settings. *WildScore* differs by grounding symbolic music analysis in online discourse, incorporating informal reasoning and context-dependent ambiguity from platforms like Reddit (Reddit, 2024).

2.2 Optical Music Recognition (OMR)

Optical Music Recognition (OMR) aims to transcribe printed or handwritten scores into machine-readable symbolic formats. Traditional systems such as Audiveris (Audiveris, 2025) and SmartScore (Musitek Corporation, 2024) focus on improving transcription accuracy under controlled input conditions. Surveys like Rebelo et al. (2012) document the progress and limitations of OMR systems, especially in their failure to handle degraded

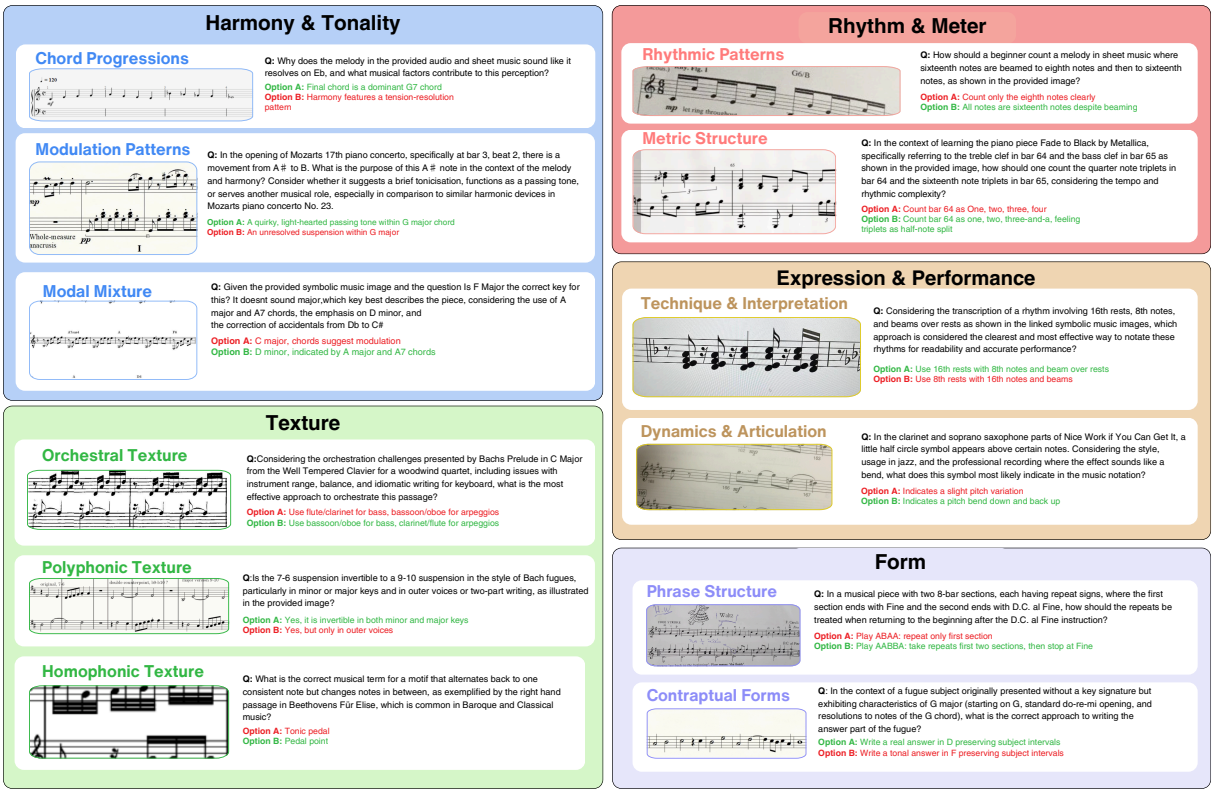


Figure 1: Example questions from our symbolic music benchmark dataset, illustrating the diversity of high-level categories and subcategories included. For each of the five core categories—Harmony & Tonality (HT), Rhythm & Meter (RM), Texture (Tx), Expression & Performance (EP), and Form (FM)—we present representative samples spanning their respective subcategories. Each panel shows a sample multiple-choice question along with corresponding answer choices, demonstrating the range and depth of musical concepts assessed in our benchmark.

or context-rich visual inputs. Recent approaches attempt deep learning-based segmentation and classification (Tuggener et al., 2018; Pecina et al., 2017), but the field still lacks benchmarks that demand semantic or contextual reasoning. *WildScore* extends OMR beyond literal transcription by introducing tasks where score fragments must be interpreted in natural language conversations. Unlike OMR, which primarily targets transcription accuracy, *WildScore* evaluates interpretive reasoning that combines visual perception of notation with higher-level musicological analysis in a QA setting.

2.3 Multimodal Reasoning with Vision-Language Models

Multimodal Large Language Models (MLLMs) like LLaVA (Liu et al., 2023a), BLIP-2 (Li et al., 2023), and Qwen-VL (Bai et al., 2023) have achieved strong performance on benchmarks such as VQAv2 (Goyal et al., 2017) and COCO (Lin et al., 2015), yet these existing benchmarks predominantly feature everyday scenes, charts, or documents and lack the formal structure and semantic density found in symbolic music notation. Unlike natural images instruction tuning (Liu et al., 2023a; Wu et al., 2025e, 2024a), music scores encode layered information through a specialized visual grammar, requiring models to integrate not just visual recognition but also domain-specific reasoning across harmony, rhythm, form, and expression. *WildScore* introduces symbolic music as a distinct and underexplored multimodal reasoning domain.

While prior music-related multimodal benchmarks focus on audio-language or audio-visual tasks (Wu et al., 2025b,c), additional recent efforts such as AIR-Bench (Yang et al., 2024), MMAU (Sakshi et al., 2024), MMAR (Ma et al., 2025), and EMOPIA (Hung et al., 2021) evaluate multimodal models in the audio channel. These are highly relevant for multimodal evaluation but do not address the complexities of symbolic *visual* music notation. By contrast, *WildScore* uniquely targets symbolic score images as a structured, visually dense modality, requiring models to parse notation and reason about harmony, rhythm, form, and expression.

Dataset	Source Type	Input Type	Multi-modal	Reasoning	Category Diversity	Real-world Content	Annotation Type	Eval. Format
MusicTheory-Bench (Czajka et al., 2024b)	Expert-curated	Theory MCQ	✗	✓	2	✗	Manual	Acc.
MAESTRO (Hawthorne et al., 2019)	Curated (Competition recs.)	Audio	✗	✗	2	✗	Automatic	F1-score
MusicNet (Thickstun et al., 2017b)	Curated (Classical recs.)	Audio	✗	✗	N/A	✗	Manual	Recall
NES-MDB (Donahue et al., 2018)	Curated (Game Audio)	Audio	✗	✗	11	✗	Automatic	N/A
MusicScore (Lin et al., 2024)	Curated (Public Scores)	Score image corpus	✓	✗	8	✗	Automatic	FID
Lakh MIDI (Raffel, 2016)	In-the-wild (Web-MIDI)	MIDI	✗	✗	N/A	✓	Automatic	N/A
WildScore	**In-the-wild (forums)**	**Musicological MCQ**	✓	✓	**5 core + 12 subcats.**	✓	**Manual + Auto-mated**	**Acc.**

Table 1: Comparison of symbolic music datasets and benchmarks. WildScore uniquely combines multimodal symbolic input, real-world musicological queries, and deep reasoning evaluation.

This positions *WildScore* as a necessary addition to the multimodal reasoning landscape, extending evaluation beyond natural images and audio into the structured world of symbolic music. Relative to theory-only question sets (knowledge recall) and OMR (perception), *WildScore* spans both knowledge-based tasks (e.g., rhythm counting) and multi-step reasoning tasks (e.g., orchestration or harmonic function in context), providing a bridge task for the reasoning community.

3 WildScore

The aim of this study is to evaluate the visual context understanding of Multimodal Large Language Models (MLLMs) for symbolic musical score as shown in Figure 2. To this end, we introduce **WildScore**. In this section, we describe the dataset details; Section 4 presents the evaluation of vision–language reasoning over symbolic musical scores.

Our dataset creation process involves two distinct phases: (1) data collection (3.1), (2) multimodal filtering (3.2). Together, these phases enforce sample relevance, symbol-image grounding, and rigorous quality control, yielding a benchmark that robustly evaluates MLLMs' visual context understanding of symbolic music.

3.1 Data Collection

This benchmark is sourced from public question posts on *r/musictheory* subreddit, covering discussions and interactions spanning over a period of ten-year period (2012–2022). This in-the-wild sourcing yielded a user-generated benchmark, with questions standardized into a canonical form for consistent evaluation while preserving their original intent. We extracted original submissions along with their corresponding first-level comments. Many submissions included embedded score images, which we extracted as the visual context for evaluation.

3.2 Multimodal Filtering

As an initial screen, we fine-tuned a YOLO (Khanam and Hussain, 2024) based detector on 215 manually annotated images using Roboflow (Dwyer et al., 2025), and then applied the detector to 4000 candidate images extracted from submissions. Each selected image was paired with the associated submission text and first-level comments. To ensure clarity and meaningful community engagement, we performed content and engagement filtering. We excluded submissions exceeding 200 words and retained only those with at least three first-level comments. This filtering pipeline resulted in a refined dataset of 807 high-

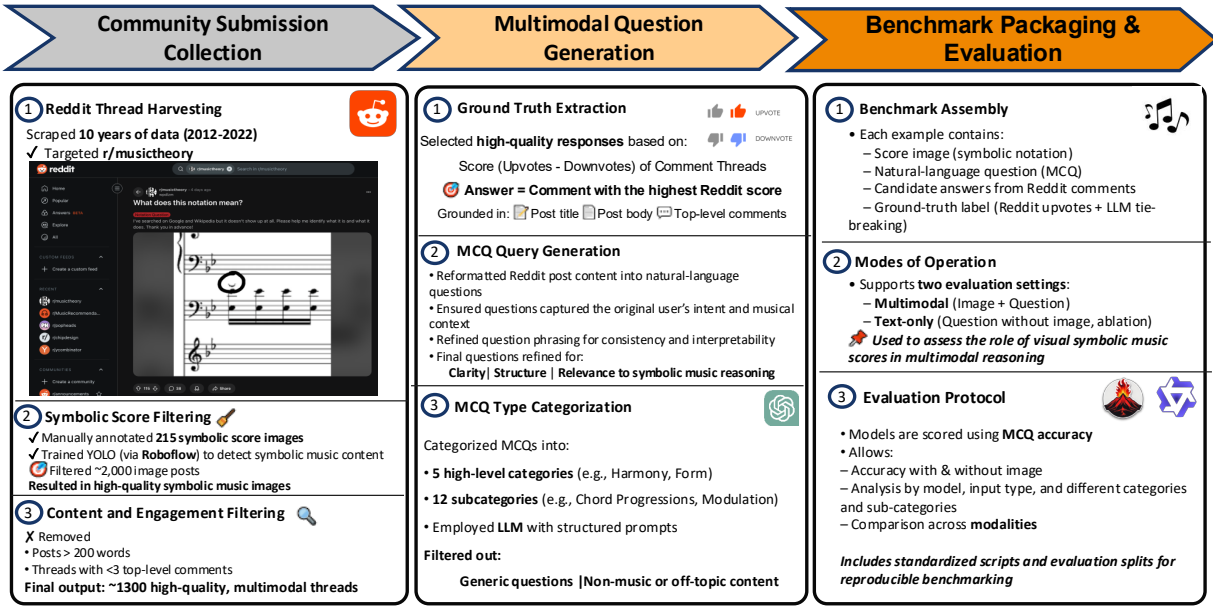


Figure 2: Overview of the dataset construction pipeline, including Reddit post collection, music entity extraction, query generation, and candidate retrieval.

quality examples.

Each dataset entry was then reformatted into multiple-choice questions (MCQs) using GPT-4.1-mini, which helped transform user queries and corresponding comments into meaningful exam-like MCQs. To establish the ground truth for each MCQ, we leveraged Reddit's engagement metrics, calculating the score as follows:

$$S = U - D$$

where S is the score, U is the number of upvotes, and D is the number of downvotes a comment has. The comment with the highest score was considered the ground truth answer. In the event of a tie, we used a language-model judge (Appendix A) to select the response best grounded in the question context. This is referred to as language-model preference, whereas the option selected according to the score S is denoted as human preference. The corresponding distributions of these preferences are presented as Annotation Preference in Table 2.

After establishing the ground truth answers, we created additional nuanced distractor options, carefully crafted with subtle distinctions from the correct responses using language model as specified in Prompt A. These options were then combined with the ground truth answers to finalize the multiple-choice benchmark dataset as shown in Figure 2.

Annotation Preference	Samples
Human preference	549
Language-model preference	258

Table 2: Distribution of WildScore questions by annotation preference.

3.3 Dataset Categorization

To support structured analysis and evaluation, we categorized our dataset into five categories as shown in Figure 3 to represent core aspects of music theory. These categories are further divided into twelve detailed subcategories as shown in Figure 4:

- **Harmony & Tonality**: Harmony concerns the progression of chords and their simultaneous combination and Tonality is the hierarchical organization of pitches around a tonal center that imparts direction and resolution (Kaliakatsos-Papakostas et al., 2025).
- **Rhythm & Meter**: The temporal aspect of music, created by the timing of musical notes and silences, establishes patterns known as rhythm. The arrangement of rhythms into regular beat patterns, frequently divided into measures, is referred to as meter (de Haas and Volk, 2016).
- **Texture**: Texture refers to the combination of melodic, harmonic, and rhythmic elements in

a composition, which might be monophonic (having only one melody) or polyphonic (having several separate lines) (Couturier et al., 2022).

- **Expression & Performance**: Expression conveys musical meaning through dynamics, articulation, phrasing, and tempo; performance is the realization of the score in sound, integrating technique and expressivity (Xia, 2016).
- **Form**: Form refers to the structure of a piece, describing the introduction, repetition, variation, and development of musical ideas (von Rütte et al., 2022).

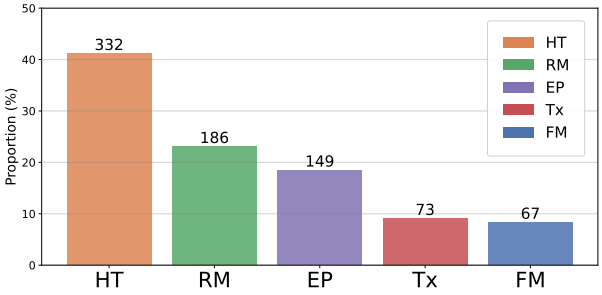


Figure 3: Distribution of symbolic music questions by high-level category. **Category abbreviations:** FM: Form, HT: Harmony & Tonality, RM: Rhythm & Meter, Tx: Texture, EP: Expression & Performance.

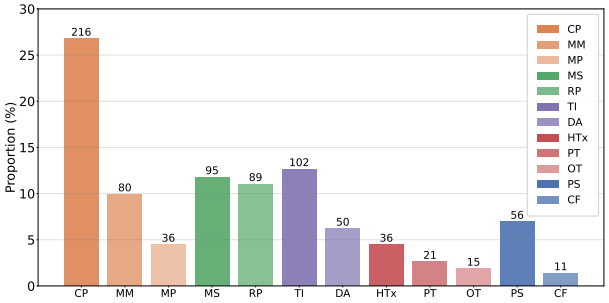


Figure 4: Distribution of symbolic music questions by subcategory. **Subcategory abbreviations:** PS: Phrase Structure, CF: Contrapuntal Forms, CP: Chord Progressions, MP: Modulation Patterns, MM: Modal Mixture, MS: Metric Structure, RP: Rhythmic Patterns, HTx: Homophonic Texture, PT: Polyphonic Texture, OT: Orchestral Texture, DA: Dynamics & Articulation, TI: Technique & Interpretation.

3.4 Dataset Overview & Statistics

The final benchmark comprises **807** items, each pairing a musical-score image with a question sourced from a Reddit submission and grounded in at least three distinct top-level comments. After manual review by three Level-3 students (Table 6), ambiguous, musically incorrect, irrelevant, or offensive items were removed. Ground-truth labels are split between **549** human-preferred items and **258** language-model-preferred items.

Difficulty stratification: We assign each question to *Easy*, *Medium*, or *Hard* using an LLM-based rubric. Specifically, we prompt GPT-4.1 with a few-shot template designed from examples curated by a Level 5 expert (criteria in Table 6) to rate the expected difficulty from the prompt. The resulting distribution is as shown in Table 7. Additional construction details, annotator instructions, and prompt templates are provided in the appendix C.

4 Experiments

We systematically evaluate several state-of-the-art MLLMs using our newly proposed symbolic music reasoning benchmark, WildScore. This evaluation examines MLLM capabilities across the five major musical categories defined by our taxonomy: Expression & Performance, Form, Harmony & Tonality, Rhythm & Meter, and Texture. We consider two evaluation settings, (1) *image+text* (symbolic score images provided) and (2) *text-only*, thereby isolating the effect of visual context and permitting direct comparison across modalities.

4.1 Evaluation Metrics

Following standard practice in multimodal reasoning benchmarks (Yu et al., 2023), we adopt accuracy as our primary metric, calculated as the percentage of correctly answered multiple-choice questions. Each question includes one correct answer, annotated based on human or language model preference as detailed in Section 3.

4.2 Quantitative Results

Across categories, GPT-4.1-mini attains the best average performance on WildScore, reaching 68.31% accuracy under the image and text setting. In the text-only setting, its accuracy declines to 65.76%, a decrease of 2.55% points, indicating a consistent benefit from visual context. Per-category accuracies for all models are reported in Table 3, with a summary visualization in Figure 5.

Performance varies significantly across categories. Notably, GPT-4.1-mini achieves the highest accuracy in *Expression & Performance* (72.12%)

and *Harmony & Tonality* (70.14%) categories, whereas it notably struggles in *Rhythm & Meter* (63.20%) and *Texture* (64.15%). This pattern aligns with our hypothesis that current MLLMs are adept at more superficial symbolic score recognition but find difficulties in tasks requiring deep symbolic abstraction and rhythmic interpretation.

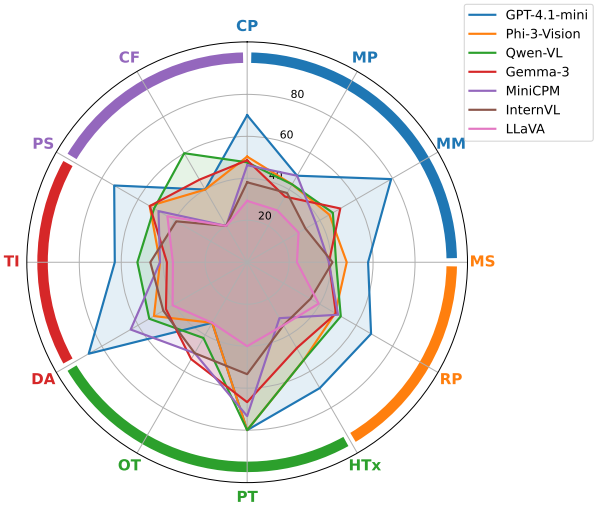


Figure 5: Per-Subcategory QA Accuracy by Vision-Enabled Model

A subcategory analysis (Table 4) reveals heterogeneous image contributions across models. For GPT-4.1-mini, accuracies peak on *Dynamics and Articulation* (87.18%) and *Modal Mixture* (79.25%) and drop on *Orchestral Texture* (33.33%) and *Contrapuntal Forms* (40.00%). Other systems show attenuated or even negative image gains in multiple subcategories. We hypothesize that this heterogeneity reflects differences in multimodal pretraining and alignment in models with stronger vision–language objectives and instruction tuning appear better grounded in symbolic notation, whereas those trained primarily on natural-image corpora or with weaker visual adapters show limited benefit from images. A Level-5 human expert was evaluated on 100 proportionally sampled questions spanning easy, medium, and hard categories, achieving an overall average accuracy of 72%. Additionally, besides *Contrapuntal Forms* most model performance *without* the image is better than random guessing, highlighting that the naturalized data used to create WildScore may not fully require perception of the scores. This makes an interesting contrast to recent synthetically difficult benchmarks that force multimodal perception to succeed (Zang et al., 2025), as such difficult benchmarks may not reflect the real distribution of questions, where perception may not always be necessary.

4.3 Limitations of Smaller Models

Among smaller MLLMs - Phi-3-Vision, Qwen-2.5-VL, Gemma-3, MiniCPM, InternVL, and LLaVA, absolute accuracies remain below GPT-4.1-mini. Within this group, Phi-3-Vision shows a small improvement with images (48.82% with image and text vs. 47.72% only with text), and Qwen-2.5-VL likewise benefits from images; Gemma-3 also shows a modest gain (46.34% vs. 44.36%). By contrast, InternVL (39.34% vs. 45.54%), MiniCPM (45.90% vs. 52.09%), and LLaVA (32.97% vs. 37.16%) are lower with images than without.

These patterns indicate that the ability to exploit symbolic score images is model-dependent. In three models: InternVL, MiniCPM, and LLaVA, the image with text setting reduces accuracy relative to only-text setting, suggesting difficulties with notation-heavy visuals and symbol prompt alignment. By contrast, Qwen-2.5-VL, Phi-3-Vision, and Gemma-3 show only modest gains from adding images. We will discuss likely failure modes: perception of basic symbols, grounding between regions of the score and the question, and higher-level reasoning over image and question understanding and potential causes in our Error Analysis (Section 4.4). We also outline directions for improvement there, including greater exposure to schematic notation during pretraining, stronger vision–language alignment for symbolic artifacts, and structure-aware encoders tailored to musical scores.

4.4 Error Analysis

We categorized failures along two axes: **perception-based errors** (reading notational symbols from the image) and **reasoning-based errors** (applying music-theory rules once symbols are correctly read). Failures that persist after successful perception are interpreted as reasoning-related failures. To evaluate perception-specific failures, we designed two diagnostic tasks: (i) a perception-only probe, and (ii) a score reconstruction on image inputs. We subsequently evaluated our best-performing model (GPT-4.1-mini) against our weakest-performing models (InternVL and LLaVA) on these tasks.

Diagnostic 1: Perception-only probe: To isolate low-level visual perception from downstream

Model	Modality	Expr. & Perf.	Form	Harmo. & Ton.	Rhythm & Meter	Texture	Average
GPT-4.1-mini (OpenAI, 2023)	w/ Image	72.12	69.57	70.14	63.20	64.15	68.31
Params: *undisclosed*	w/o Image	67.31	71.74	64.25	67.20	60.38	65.76
Qwen-2.5-VL (Bai et al., 2023)	w/ Image	52.88	52.17	47.06	47.20	58.49	49.73
Params: *8.29B*	w/o Image	51.92	52.17	46.15	46.40	60.38	49.18
Phi-3-Vision (Abdin et al., 2024)	w/ Image	45.19	52.17	48.42	48.00	56.60	48.82
Params: *4.15B*	w/o Image	46.15	45.65	47.51	47.20	54.72	47.72
Gemma-3 (Team et al., 2025)	w/ Image	40.27	52.24	47.89	43.55	53.42	46.34
Params: *4.3B*	w/o Image	46.31	49.25	42.47	42.47	49.32	44.36
MiniCPM (Hu et al., 2024)	w/ Image	50.00	45.65	44.34	44.80	47.17	45.90
Params: *3.43B*	w/o Image	57.69	54.35	49.32	48.80	58.49	52.09
InternVL (Chen et al., 2023)	w/ Image	46.15	36.96	36.65	37.60	43.40	39.34
Params: *9.14B*	w/o Image	52.88	45.65	40.27	44.00	56.60	45.54
LLaVA (Liu et al., 2023b)	w/ Image	37.50	41.30	28.96	32.00	35.85	32.97
Params: *7.06B*	w/o Image	40.38	50.00	33.03	35.20	41.51	37.16

Table 3: Per-category accuracy (%) by model and input modality. Model sizes (Params) are shown under model names.

Model	Modality	Harmony & Tonality			Rhythm & Meter		Texture			Express. & Perfor.		Form		Average
		CP	MP	MM	MS	RP	HTx	PT	OT	DA	TI	PS	CF	
GPT-4.1-mini	w/ Image	70.07	47.62	79.25	57.63	68.18	69.23	80.00	33.33	87.18	63.08	73.17	40.00	68.31
	w/o Image	63.95	52.38	69.81	61.02	72.73	57.69	80.00	41.67	84.62	56.92	73.17	60.00	65.76
Qwen-VL	w/ Image	47.62	42.86	47.17	42.37	51.52	53.85	80.00	41.67	53.85	52.31	51.22	60.00	49.73
	w/o Image	46.26	42.86	47.17	40.68	51.52	57.69	80.00	41.67	48.72	53.85	51.22	60.00	49.18
Phi-3-Vision	w/ Image	50.34	42.86	45.28	47.46	48.48	53.85	80.00	33.33	51.28	41.54	53.66	40.00	48.82
	w/o Image	50.34	57.14	35.85	49.15	45.45	50.00	80.00	33.33	48.72	44.62	48.78	20.00	47.72
Gemma-3	w/ Image	48.61	36.11	51.25	38.95	48.86	47.22	66.67	53.33	44.68	38.24	53.57	45.45	46.34
	w/o Image	43.52	30.56	45.00	37.89	47.73	44.44	57.14	53.33	57.45	41.18	51.79	36.36	44.36
MiniCPM	w/ Image	46.26	47.62	37.74	38.98	50.00	30.77	73.33	50.00	64.10	41.54	48.78	20.00	45.90
	w/o Image	52.38	42.86	43.40	44.07	53.03	53.85	60.00	66.67	64.10	53.85	53.66	60.00	52.09
InternVL	w/ Image	38.10	38.10	32.08	40.68	34.85	34.62	53.33	50.00	46.15	46.15	39.02	20.00	39.34
	w/o Image	42.18	23.81	41.51	54.24	34.85	46.15	80.00	50.00	46.15	56.92	48.78	20.00	45.54
LLava	w/ Image	29.25	28.57	28.30	23.73	39.39	34.62	40.00	33.33	41.03	35.38	43.90	20.00	32.97
	w/o Image	30.61	33.33	39.62	25.42	43.94	38.46	46.67	41.67	35.90	43.08	51.22	40.00	37.16

Table 4: Per-subcategory accuracy (%) by model and input modality, with subcategories grouped by category.

reasoning, we posed straightforward factual queries (e.g., clef identification, symbol counts) on 50 symbolic-score images from our benchmark. The items were handcrafted by two Level 3 (Table 6) human experts to avoid higher-level inference. Accuracy on this probe is shown in Table 5. GPT-4.1-mini correctly perceived relevant symbols in 52% of cases, whereas InternVL and LLaVA reached 38% and 26%, respectively. These results indicate that a substantial portion of smaller-model errors originate at the perception stage rather than from subsequent reasoning.

Table 5: Perception-only probe (accuracy %) on 50 symbolic-score images. Higher is better.

Model	Accuracy (%)
GPT-4.1-mini	52.0
InternVL	38.0
LLaVA	26.0

Diagnostic 2: Score reconstruction from images: We further examined end-to-end symbol extraction by asking models to produce ABC notation directly from score images. We evaluate outputs for syntactic validity and bar level faithfulness (qualitative summaries in Table 8). InternVL and LLaVA frequently generated invalid or degenerate sequences (e.g., looping a single chord), while GPT-4.1-mini produced valid ABC notations for simpler, single-staff excerpts but degraded on longer or denser passages, often with omissions or repeated bars. These outcomes point to limits in sustained symbolic tracking rather than purely textual reasoning.

Across both diagnostics, smaller models struggle to accurately read notation reliably, and failures in perception propagate to reasoning. GPT-4.1-mini shows stronger symbol reading and can reconstruct short excerpts, but still falters on longer contexts, indicating residual limits in reasoning over extended structure. These findings align with the heterogeneous image effects observed and suggest that improving pretraining on notation-heavy corpora and strengthening vision–to–symbol extraction are prerequisites for consistent gains on symbolic music reasoning.

5 Conclusion

In this work, we have introduced WildScore, a benchmark designed to evaluate the capabilities of Multimodal Large Language Models (MLLMs) in symbolic music reasoning with visual context. WildScore captures the richness and diversity of real-world musicological conversation by utilizing real musical scores in conjunction with community-sourced questions and answers from Reddit. Our systematic taxonomy, encompassing broad musical categories and detailed subcategories, facilitates nuanced evaluation and identification of model strengths and limitations.

Empirical results indicate that while current state-of-the-art MLLMs exhibit substantial promise, particularly in tasks involving surface-level recognition and straightforward analysis, they continue to struggle significantly with deep symbolic abstraction, rhythmic complexity, and orchestration intricacies especially when presented as an image. Significant differences in performance demonstrated by popular multimodal large language models between text-only and visual inputs highlight how important visual context is for precise musicological interpretation.

Furthermore, our analysis highlights the substantial limitations of smaller-scale models, suggesting that significant advancements in symbolic music understanding remain necessary. WildScore thus not only fills a crucial gap in multimodal music reasoning benchmarks but also sets a clear trajectory for future research efforts aimed at enhancing the depth and nuance of symbolic musical comprehension in multimodal frameworks.

Limitations

Reddit's ranking mechanisms often favor mainstream topics, which may distort the visibility of niche symbolic music practices and reinforce dominant stylistic norms. Despite filtering, some comments may contain informal or toxic language. Symbolic music discussions may also be misinformed or lack technical rigor, which affects their utility for modeling.

Ethical considerations

Data Collection and Anonymization This dataset is constructed from publicly available Reddit posts, collected via the official Reddit API in compliance with the platform's Content Policy and Terms of Use. All usernames, IDs, and personal metadata have been removed to ensure anonymity. Although Reddit is a public forum, we acknowledge that users may not anticipate their contributions being used for research, particularly in academic or computational contexts.

Use and Licensing The dataset is released under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). It is intended strictly for non-commercial research. We highly urge researchers to consider the ethical implications of modelling public discourse, especially in creative and culturally sensitive domains like symbolic music, where interpretations may carry stylistic or cultural assumptions.

Acknowledgments

This work was partially supported by the U.S. National Science Foundation under Grant IIS-2432486.

LLM Usage: We used large language models solely for grammar refinement and minor wording edits in drafting parts of this paper.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Audiveris. 2025. Audiveris. `https://github.com/Audiveris/audiveris`. Version 5.6.2. Released 2025-07-18. Accessed: 2025-08-31.

Jinze Bai and 1 others. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Kang Chen and Xiangqian Wu. 2024. Vtqa: Visual text question answering via entity alignment and cross-media reasoning. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 27218–27227.

Zhe Chen and 1 others. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Louis Couturier, Louis Bigo, Florence Levé, and Markus Neuwirth. 2022. A dataset of symbolic texture annotations in mozart piano sonatas. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, pages 509–516.

Lukasz Czajka, Malihe Alikhani, Patrick Verga, and Yonatan Belinkov. 2024a. Musical understanding benchmark: A challenge for large language models. *arXiv preprint arXiv:2410.02084*.

Lukasz Czajka, Malihe Alikhani, Patrick Verga, and Yonatan Belinkov. 2024b. Musictheory-bench. ArXiv preprint arXiv:2410.02084.

W. Bas de Haas and Anja Volk. 2016. Meter detection in symbolic music using inner metric analysis. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 441–447.

Chris Donahue, Zachary C. Lipton, and Julian McAuley. 2018. The nes music database: A multi-instrumental dataset with expressive performance attributes. In *ISMIR*.

Brad Dwyer, Joseph Nelson, Tyler Hansen, and 1 others. 2025. Roboflow (version 1.0). Computer vision software.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse H. Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shengding Hu and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, and 1 others. 2025. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*.

Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. 2021. Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *Preprint*, arXiv:2108.01374.

Md Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Md Azam Hossain. 2025. Visual robustness benchmark for visual question answering (vqa). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6623–6633. IEEE.

Maximos Kaliakatsos-Papakostas, Dimos Makris, Konstantinos Soiledis, Konstantinos-Theodoros Tsamis, Vassilis Katsouros, and Emilios Cambouropoulos. 2025. Harmonytok: Comparing methods for harmony tokenization for machine learning. *Preprints*.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Junnan Li, Dongxu Li, and 1 others. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Yuheng Lin, Zheqi Dai, and Qiuqiang Kong. 2024. Musicscore: A dataset for music score modeling and generation. *arXiv preprint arXiv:2406.11462*.

Haotian Liu, Chunyuan Zhang, and 1 others. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Haotian Liu and 1 others. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. Mmar: A challenging benchmark for deep

reasoning in speech, audio, music, and their mix. *Preprint*, arXiv:2505.13032.

Musitek Corporation. 2024. Smartscore 64 ne pro edition. https://www.musitek.com/smartscore64-pro.html. Accessed: 2025-08-31.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pavel Pecina and 1 others. 2017. In search of a dataset for handwritten optical music recognition: Introducing muscima++. *arXiv preprint arXiv:1703.04824*.

Colin Raffel. 2016. Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching. PhD dissertation, Columbia University.

Ana Rebelo, João Pimentel, Jaime Cardoso, and 1 others. 2012. Optical music recognition: state-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, 1(3):173–190.

Reddit. 2024. Reddit. https://www.reddit.com/. Accessed: 2025-08-31.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

Rohan Surana, Aakash Varshney, and Vishnu Pendyala. 2022. Deep learning for conversions between melodic frameworks of indian classical music. In *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, pages 1–12, Singapore. Springer Nature Singapore.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade. 2017a. Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*.

John Thickstun, Zaid Harchaoui, and Sham M Kakade. 2017b. Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*.

Lukas Tuggener, Ismail Elezi, Jurgen Schmidhuber, Marcello Pelillo, and Thilo Stadelmann. 2018. Deepscores-a dataset for segmentation, detection and classification of tiny objects. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3704–3709. IEEE.

Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. 2022. Figaro: Generating symbolic music with fine-grained artistic control. *arXiv preprint arXiv:2201.10936*.

Ruoyu Wang, Tong Yu, Junda Wu, Yao Liu, Julian McAuley, and Lina Yao. 2025. Weakly-supervised vlm-guided partial contrastive learning for visual language navigation. *arXiv preprint arXiv:2506.15757*.

Junda Wu, Jessica Echterhoff, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. 2025a. Pdb-eval: An evaluation of large multimodal models for description and explanation of personalized driving behavior. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 242–248. IEEE.

Junda Wu, Warren Li, Zachary Novack, Amit Namburi, Carol Chen, and Julian McAuley. 2025b. Collap: Contrastive long-form language-audio pretraining with musical temporal structure augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. 2024a. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, and 1 others. 2024b. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*.

Junda Wu, Zachary Novack, Amit Namburi, Hao-Wen Dong, Carol Chen, Jiaheng Dai, and Julian McAuley. 2025c. Futga-mir: Enhancing fine-grained and temporally-aware music understanding with music information retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, and 1 others. 2025d. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, and 1 others. 2025e. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, and 1 others. 2024c. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Gus Xia. 2016. *Expressive Collaborative Music Performance via Machine Learning*. Ph.D. thesis, Carnegie Mellon University.

Weihan Xu, Julian McAuley, Taylor Berg-Kirkpatrick, Shlomo Dubnov, and Hao-Wen Dong. 2024. Generating symbolic music from natural language prompts using an llm-enhanced dataset. *Preprint*, arXiv:2410.02084.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, and 1 others. 2024. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *Preprint*, arXiv:2402.07729.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, and 1 others. 2024. Chatmusician: Understanding and generating music intrinsically with llms. *arXiv preprint arXiv:2402.16153*.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *arXiv preprint arXiv:2504.00369*.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. 2024. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*.

A Prompt Templates

Prompt 1: Multimodal Answer Selection (With Image)

System Prompt:
You are an expert in symbolic music-score question answering. You will be provided with an image of a musical excerpt, a question about it, and several labeled options. Analyze the image and text, then choose the correct answer. Respond with **ONLY** the option letter.

User Prompt:

```
<image>

Question: Which measure best represents the 6/8 time signature?

Options:
A. Grouped in two dotted-quarter notes
B. Grouped as three quarter notes
```

Prompt 2: Text-Only Answer Selection

System Prompt:
You are an expert in symbolic music-score question answering. You will be provided with a question about a musical excerpt and several labeled options. Choose the correct answer based solely on the text. Respond with **ONLY** the option letter.

User Prompt:

```
Question: Which measure best represents the 6/8 time signature?

Options:
A. Grouped in two dotted-quarter notes
B. Grouped as three quarter notes
```

Prompt 3: Distractor Generation

System Prompt:
You are a musicology professor preparing multiple-choice questions for an upcoming exam. You are given a music-related question and one correct option. Generate nuanced distractor options with subtle differences from the correct answer.

Guidelines:
- Generate up to three distractors (fewer is fine).
- They must all be plausible yet incorrect.
- Keep them concise (5–10 words).

Return ONLY valid JSON in the form: `{"Option A": "...", "Option B": "...", ...}`

User Prompt:

```
"Title": <title_of_reddit_submission>
"Question": <reformatted_question>
"Correct Option": <decided_ground_truth_answer>
```

Prompt 4: Ground-Truth Selection (Text-Only)

System Prompt:
You are an expert in symbolic music-score question answering. You will be provided with a question about a musical excerpt and several labeled options. Choose the correct answer based on the text. Respond with **ONLY** the option letter.

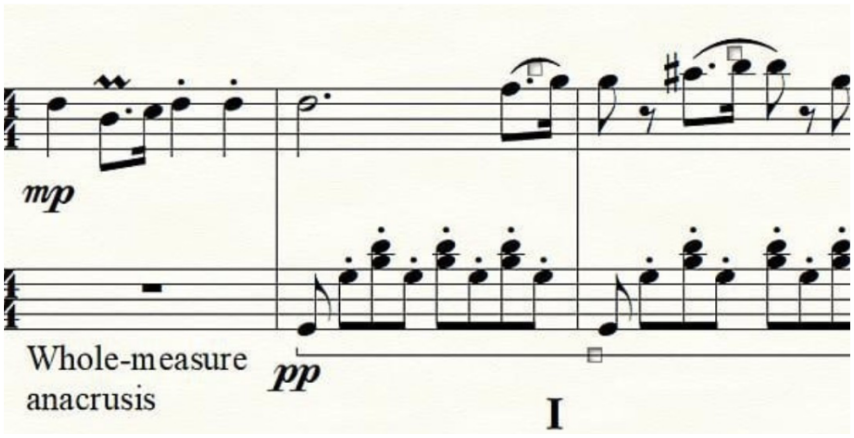
User Prompt:

```
Question: {{QUESTION_PLACEHOLDER}}

Options:
{{OPTIONS_PLACEHOLDER}}
```

B Illustrative Items

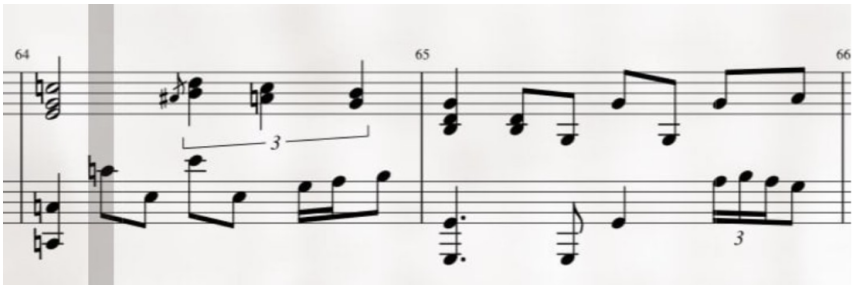
Harmony & Tonality



QUESTION ht2: In the opening of Mozart's 17th piano concerto, specifically at bar 3, beat 2, there is a movement from A# to B. What is the purpose of this A# note in the context of the melody and harmony? Consider whether it suggests a brief tonicisation, functions as a passing tone, or serves another musical role, especially in comparison to similar harmonic devices in Mozart's piano concerto No. 23.

A. A quirky, light-hearted passing tone within G major chord
B. An unresolved suspension within G major
C. A leading tone preparing for modulation
D. A dominant note resolving to G major

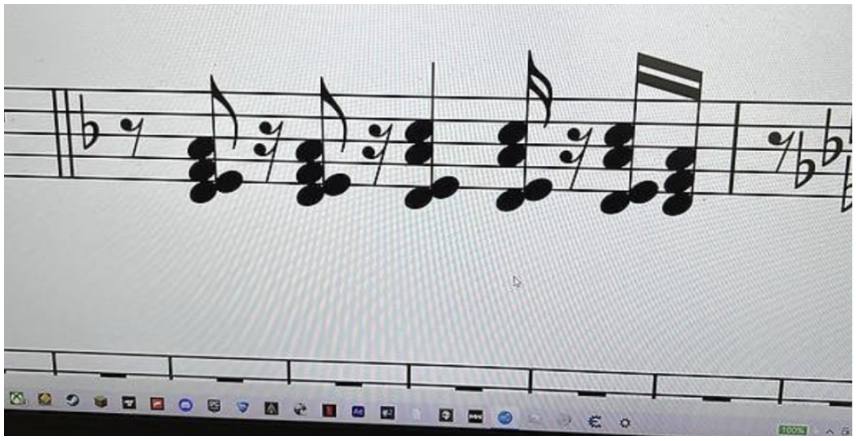
Rhythm & Meter



In the context of learning the piano piece *Fade to Black* by Metallica, specifically referring to the treble clef in bar 64 and the bass clef in bar 65 as shown in the provided image, how should one count the quarter-note triplets in bar 64 and the sixteenth-note triplets in bar 65, considering the tempo and rhythmic complexity?

A. Count bar 64 as one, two, three, four
B. Count bar 64 as one-and-two-and, steady quarter
C. Count bar 64 as one, two, three-and-a, feeling triplets as half-note split
D. Count bar 64 as one-and-two, triplet feel

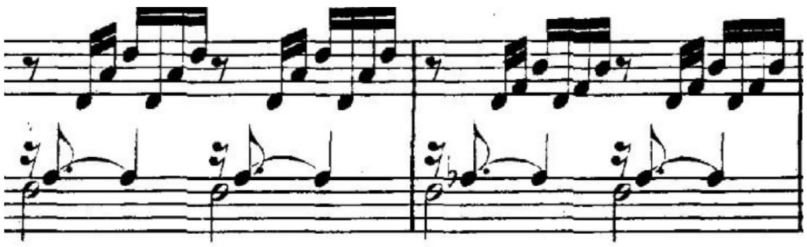
Expression & Performance



Considering the transcription of a rhythm involving 16th rests, 8th notes, and beams over rests as shown in the linked symbolic music images, which approach is considered the clearest and most effective way to notate these rhythms for readability and accurate performance?

A. Use quarter notes and beams over rests
B. Use 16th notes only with no rests
C. Use 16th rests with 8th notes and beam over rests
D. Use 8th rests with 16th notes and beams

Texture



QUESTION: Considering the orchestration challenges presented by Bach's *Prelude in C Major* from the *Well-Tempered Clavier* for a woodwind quartet, including issues with instrument range, balance, and idiomatic writing for keyboard, what is the most effective approach to orchestrate this passage?

A. Use flute/clarinet for bass, bassoon/oboe for arpeggios
B. Use clarinet for melody, bassoon for counterpoint
C. Use bassoon/oboe for bass, clarinet/flute for arpeggios
D. Use oboe/bassoon for harmony, flute/clarinet for melody

C Human Expertise Criteria

Table 6: Assessment levels for human expertise.

Level	Description
1	Rarely listens to music.
2	No music-theory knowledge, but can distinguish genres and has preferred styles.
3	Basic knowledge of playing an instrument or music theory.
4	No formal training; self-taught aspects of music theory.
5	Completed academic coursework in music theory.

Full instruction text shown to annotators. You are asked to review candidate MCQ items derived from Reddit submissions that include musical-score images. For each item: (i) check that the question is musically correct and unambiguous; (ii) verify that the answer options are relevant to the question; (iii) delete any options or posts you judge irrelevant or offensive; (iv) flag any ambiguous or musically incorrect items for exclusion. If you encounter potentially offensive material, do not continue with that item—remove/flag it and proceed to the next one. Do not record or transcribe any personal identifying information (PII) that might appear in posts or images.

Recruitment and compensation. Annotators were Level-3 students at a U.S. university and received course credit. Participation was voluntary; no monetary payments were provided.

Consent and data provenance. By opting into the course-credit activity, annotators consented to their contributions being used for research. Reddit content was obtained from publicly available posts via the official API; usernames and direct identifiers were removed, and use followed the platform's terms.

Ethics determination. This project analyzes public data and involves low-risk student annotation without collection of PII; it was determined that formal IRB review was not required.

Demographics. No annotator demographic data were collected.

Table 7: Distribution of datapoints by LLM-assigned difficulty tier.

Tier	Count
Easy	191
Medium	573
Hard	43
Total	807

Table 8: ABC reconstruction from images: qualitative outcomes.

Model	Observed outcome
GPT-4.1-mini	Often produces valid, faithful ABC; reliability drops in extended sequences, with omissions or repeated bars.
InternVL	Frequently yields invalid or incorrect ABC; many degenerate sequences.
LLaVA	Predominantly generates degenerate loops and invalid ABC.

D Bias Check for Judge/Formatting

To assess whether using GPT-4.1-mini in the pipeline could bias evaluation, we compared GPT-4.1 and GPT-4.1-mini on a random 50-item subset drawn from WildScore under the same protocol.GPT-4.1 secured 58 % accuracy while GPT-4.1-mini only secured 50 % accuracy as seen in Table 9.

Table 9: Subset comparison (50 items) probing potential bias from using GPT-4.1-mini in data construction.

Model	Accuracy (%)	n
GPT-4.1	58	50
GPT-4.1-mini	50	50