

← Go to **ACM SIGIR 2026 Short Papers Track** homepage (/group?id=ACM.org/SIGIR/2026/Short\_Papers\_Track)

# CSyMR: Benchmarking Compositional Music Information Retrieval in Symbolic Music Reasoning



*Boyang Wang* (/profile?id=~Boyang\_Wang12),  
*Yash Vishe* (/profile?id=~Yash\_Vishe1), *Xin Xu* (/profile?id=~Xin\_Xu8),  
*Zachary Novack* (/profile?id=~Zachary\_Novack1),  
*Xunyi Jiang* (/profile?id=~Xunyi\_Jiang2),  
*Julian McAuley* (/profile?id=~Julian\_McAuley1),  
*Junda Wu* (/profile?id=~Junda\_Wu1)

Published: 02 Apr 2026, Last Modified: 02 Apr 2026 SIGIR 2026 short papers  
 Short Papers Track, Area Chairs, Reviewers, Authors Revisions (/revisions?id=ygsP0Jyskr)  
 BibTeX CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

**Confirm That All Author Profiles Are Complete:** The authors confirm that they have read and understood this passage from the Call for Papers, and that all authors have completed their profile accordingly

**Common Submission Requirements:** The authors confirm that they have read and understood requirements in the Call for Papers, and that their submission complies with it

**Keywords:** symbolic music, music analysis, music information retrieval, tool-augmented agent, music reasoning benchmark

## Abstract:

Natural language information needs over symbolic music scores rarely reduce to a single step lookup. Many queries require compositional Music Information Retrieval (MIR) that extracts multiple pieces of evidence from structured notation and aggregates them to answer the question. This setting remains challenging for Large Language Models due to the mismatch between natural language intents and symbolic representations, as well as the difficulty of reliably handling long structured contexts. Existing benchmarks only partially capture these retrieval demands, often emphasizing isolated theoretical knowledge or simplified settings. We introduce CSyMR-Bench, a benchmark for compositional MIR in symbolic music reasoning grounded in authentic user scenarios. It contains 126 multiple choice questions curated from community discussions and professional examinations, where each item requires chaining multiple atomic analyses over a score to derive implicit musical evidence. To support diagnosis, we provide a taxonomy with six query intent categories and six analytical dimension tags. We further propose a tool-augmented retrieval and reasoning framework that integrates a ReAct-style controller with deterministic symbolic analysis operators built with music21. Experiments across prompting baselines and agent variants show that tool-grounded compositional retrieval consistently outperforms Large Language Model-only approaches, yielding 5-7% absolute accuracy gains, with the largest improvements on analysis-heavy categories.

**Submission Number:** 351

Filter by reply type...

Filter by author...

Search keywords...

Sort: Newest First

5 / 5 replies shown




Everyone	Program Chairs	Submission351 Area...	Submission351...
----------	----------------	-----------------------	------------------

Submission351 Authors	Submission351...	Submission351...	Submission351...
-----------------------	------------------	------------------	------------------

Submission351...	✕
------------------	---

Add: **Withdrawal**

## Paper Decision

Decision by Program Chairs  31 Mar 2026, 16:23 (modified: 02 Apr 2026, 07:17)
 Program Chairs, Area Chairs, Reviewers, Authors  Revisions (/revisions?id=BqCj4DCiww)
**Decision:** Accept

## Meta Review of Submission351 by Area Chair qXa2

Meta Review by Area Chair qXa2  19 Mar 2026, 10:13 (modified: 02 Apr 2026, 07:28)
 Area Chairs, Authors, Reviewers Submitted, Program Chairs  Revisions (/revisions?id=VYQlxbi8vb)

### Metareview:

This is the meta-review by the SPC assigned to this paper.

This paper introduces CSyMR-bench, a new benchmark for compositional symbolic music IR, motivated by real-word questions and with a taxonomy of query intents. Reviewers agree that this paper fills a meaningful gap, and that the agent-based reasoning in music21 operators is a clear improvement over LLM-only prompting baselines. The identified limitations concern the size of the benchmark, the use of a single backbone, and somewhat unclear description of some aspects related to the dataset construction process, such as the distractors. The paper is not 100% clear on this, but it looks like the dataset will be publicly released. We hope that this feedback is useful to make this an even stronger contribution.

**Recommendation:** Accept**Confidence:** 3: I am fairly confident that the evaluation is correct

## Review of CSyMR: Benchmarking Compositional Music Information Retrieval in Symbolic Music Reasoning

Official Review by Reviewer Frp7  17 Mar 2026, 21:35 (modified: 02 Apr 2026, 07:33)
 Program Chairs, Area Chairs, Reviewers Submitted, Reviewer Frp7, Authors

 Revisions (/revisions?id=fSkPoiGzRt)
**Relevance To Conference:** Music Information Retrieval is a domain-specific extension to the main themes of SIGIR**Relevance Score:** 1 good**Novelty:** 1 good**Technical Soundness:** 1 good**Quality Of Presentation:** 1 good

### Strengths:

1. Novel and well-motivated benchmark. CSyMR-Bench fills a genuine gap by focusing on compositional retrieval over symbolic music scores, distinct from audio-based benchmarks like CMI-Bench (<https://arxiv.org/html/2506.12285v1> (<https://arxiv.org/html/2506.12285v1>)) or image-based ones like WildScore (<https://arxiv.org/abs/2509.04744> (<https://arxiv.org/abs/2509.04744>)). Grounding questions in real community discussions (r/musictheory) and professional exams ensures ecological validity.
2. Rich diagnostic taxonomy. The six query intent categories and six analytical dimension tags enable fine-grained analysis of model capabilities, going beyond aggregate accuracy. This is a strength for future research on targeted improvements.

3. Tool-augmented agent design is principled. Using deterministic music21 operators as structured retrieval functions is a well-grounded approach. The context isolation strategy (hiding raw code from the reasoning agent) is a thoughtful engineering choice that keeps the LLM focused on evidence aggregation. This aligns with the broader finding from "Can LLMs 'Reason' in Music?" (<https://www.researchgate.net/publication/382738860>) that current LLMs exhibit poor performance in song-level multi-step music reasoning.
4. Illuminating case study. Figure 4 provides a compelling walkthrough showing how CoT hallucinates pitches while the tool-augmented agent recovers via iterative deterministic analysis -- effectively demonstrating the value proposition of the approach.

#### Weaknesses:

1. Small benchmark size limits statistical power. With only 126 questions, subgroup analyses (e.g., per-category in Table 1) have very small sample sizes (some categories have only 14-16 items). This makes it difficult to draw reliable conclusions about per-category performance differences, and the results may not generalize.
2. Single backbone for agent experiments. All prompting baselines and agent experiments use GPT-4.1-mini. While Table 2 shows zero-shot results for multiple models, the key finding (tool augmentation helps) is only validated with one backbone. Would smaller or open-source models benefit equally from the tool-augmented framework?
3. No error analysis of the tool-augmented agent. The paper shows where tools help but does not analyze failure modes. When does the agent call the wrong tool? When does correct tool output still lead to wrong answers? This analysis would be valuable for understanding the framework's limitations.
4. OMR conversion introduces unquantified noise. The pipeline converts sheet music images to Humdrum kern via OMR software, but OMR is known to be error-prone. The paper does not report OMR accuracy or discuss how conversion errors might affect benchmark difficulty or model performance.

**Overall Recommendation:** CSyMR-Bench is a well-designed benchmark for an underexplored problem -- compositional MIR over symbolic scores. The tool-augmented agent framework demonstrates clear benefits over LLM-only baselines. However, the small benchmark size, single-backbone evaluation, and lack of error analysis limit the strength of the conclusions.

**Overall Recommendation Score:** 1 weak accept

**Nominate For Best Paper:** 0 no

**Reviewer Certification:** 1 yes

**Reviewer Confidence:** I am confident, but not absolutely certain that my evaluation is correct.

## Multiple choice music QA benchmark, and associated baseline using score analysis tool in a RAG-like manner

Official Review by Reviewer Goso  14 Mar 2026, 14:28 (modified: 02 Apr 2026, 07:34)

 Program Chairs, Area Chairs, Reviewers Submitted, Reviewer Goso, Authors

 Revisions (/revisions?id=VyLtDyOvW9)

**Relevance To Conference:** RAG/LLM-based QA is a common IR topic thesedays.

**Relevance Score:** 2 excellent

**Novelty:** 1 good

**Technical Soundness:** 1 good

**Quality Of Presentation:** 1 good

#### Strengths:

1. Interesting to see a score analysis tool (Music2, 2010) being used to retrieve pertinent score passages before selecting from 4 multiple choice answers for each question.
2. A new music multiple-choice QA dataset is introduced, collected from community forums and exams. The nature of the task is well illustrated (e.g., showing questions, and an associated Humrum kern excerpt), and the types and frequency of different questions types is concrete (and roughly 'multi-hop', i.e., requiring multiple score excerpts to be located and information from these logically combined to answer a question). The data is small, but carefully constructed, and so I think it is likely to have real value in developing related systems, and possibly help seed the creation of larger data sets of this type.

3. Fig 4 illustrates model execution nicely (I do wonder if there is a way to make this more compact, maybe by hiding some steps -- but I admit it is hard to choose what to leave out without making things less clear).

**Weaknesses:**

1. Use of the word 'significant,' but there are no statistical hypothesis tests.
2. (minor) Fig 4 helps, but it would be good to clarify the precise nature of the (e)vidence, which I initially assumed were excerpts, but seem to instead be specific analysis outputs from Music21 that a transformer/LLM has converted to prose (is that correct?).

**Overall Recommendation:** I think this paper is an interesting contribution for a small paper, and I recommend that the paper be accepted.

The work is preliminary, but carefully considered and well-described; I suspect that others will find this paper useful, both for developing similar (possibly larger) datasets and in exploring tool usage with LLMs for music QA.

This is not a weakness in my mind (because the work is preliminary), but its worth noting that the community-derived questions appear to be substantially easier to solve, and the model proposed with the music21 'tool' usage is actually outperformed on the harder exam questions by few-shot prompting of GPT (which to be honest, surprised me). In future work that may be worth carefully digging into.

Additional Notes:

\* Table 1 appears after Table 2.

**Overall Recommendation Score:** 1 weak accept

**Nominate For Best Paper:** 0 no

**Reviewer Certification:** 1 yes

**Reviewer Confidence:** I am confident, but not absolutely certain that my evaluation is correct.

## Interesting benchmark for compositional music retrieval but limited novelty and insufficient experimental rigor

Official Review by Reviewer uvX5  08 Mar 2026, 04:18 (modified: 02 Apr 2026, 07:34)

 Program Chairs, Area Chairs, Reviewers Submitted, Reviewer uvX5, Authors

 Revisions (/revisions?id=dHX3qYeyxV)

**Relevance To Conference:** The paper is clearly related to Information Retrieval, particularly domain-specific retrieval over structured symbolic music scores. It frames the task as Compositional MIR where answers require multi-step evidence extraction from structured notation rather than direct text retrieval.

**Relevance Score:** 1 good

**Novelty:** 0 fair

**Technical Soundness:** 0 fair

**Quality Of Presentation:** 1 good

**Strengths:**

1. The paper frames symbolic music reasoning as a compositional retrieval problem requiring multiple evidence extraction operations from structured scores. This perspective is conceptually appealing and could broaden IR research into new structured domains.
2. The dataset introduces query intent categories and analytical dimensions, enabling more granular evaluation of reasoning capabilities across different musical analysis tasks. This is a useful diagnostic design for benchmark construction.
3. The integration of deterministic music analysis tools (via music21) with a ReAct-style controller illustrates how grounding reasoning in verifiable symbolic operators can improve accuracy over purely parametric LLM approaches.
4. The detailed example comparing Chain-of-Thought reasoning with tool-augmented reasoning effectively demonstrates how hallucination can occur in LLM-only reasoning pipelines.

**Weaknesses:**

1. The benchmark construction pipeline uses GPT-4o-mini to synthesize distractor answers. This approach may introduce several issues: the generated distractors could be overly simplistic or unrealistic, and models may

exploit artifacts from LLM-generated options rather than performing genuine reasoning over the symbolic score. Furthermore, the paper does not describe any validation procedure to assess the quality of the generated distractors. Without expert verification, this may introduce bias into the evaluation.

2. The paper does not report important aspects of the dataset annotation process, including annotation guidelines, the number of annotators involved, inter-annotator agreement, or expert validation procedures. For benchmark datasets, such information is typically necessary to ensure the reliability and correctness of the ground-truth labels.
3. The proposed framework primarily combines an existing ReAct reasoning framework with deterministic symbolic analysis tools, which closely resembles many recent tool-augmented LLM systems. The algorithmic novelty beyond the dataset is limited.
4. The paper does not mention releasing the benchmark dataset or the code used in the experiments. Given that the dataset is the main contribution of the work, providing an anonymous repository would significantly improve reproducibility and enable the research community to build upon this benchmark.

**Overall Recommendation:** This paper introduces CSyMR-Bench, a benchmark for compositional music information retrieval over symbolic scores, and proposes a tool-augmented reasoning framework that integrates deterministic music21 operators with a ReAct-style controller. Experimental results indicate that tool-grounded reasoning improves accuracy compared to prompting-based baselines.

While framing symbolic music reasoning as a compositional retrieval task is an interesting idea, the overall contribution remains somewhat limited. The proposed framework largely follows existing tool-augmented LLM paradigms, offering limited methodological novelty beyond the dataset itself. In addition, the benchmark construction process raises concerns: distractor answers are generated using GPT-4o-mini, but the paper does not describe any validation procedure to ensure their quality, which may introduce evaluation bias. The dataset creation process also lacks details on annotation guidelines, annotator involvement, or inter-annotator agreement, making it difficult to assess label reliability.

**Overall Recommendation Score:** 0 borderline

**Nominate For Best Paper:** 0 no

**Reviewer Certification:** 1 yes

**Reviewer Confidence:** I am fairly confident that my evaluation is correct, but I have significant doubts.

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[News \(/group?id=OpenReview.net/News&referrer=\[Homepage\]\(/\)\)](/group?id=OpenReview.net/News&referrer=[Homepage](/))

[FAQ \(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

**Donate (/donate)**

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors). © 2026 OpenReview