NOTA: Multimodal Music Notation Understanding for Visual Large Language Model

Anonymous ACL submission

Abstract

Symbolic music is represented in two distinct forms: two-dimensional, visually intuitive score images, and one-dimensional, standardized text annotation sequences. While large language models have shown extraordinary potential in music, current research has primarily focused on unimodal symbol sequence text. Existing general-domain visual language models still lack the ability of music notation understanding. Recognizing this gap, we propose NOTA, the first large-scale comprehensive multimodal music notation dataset. It consists of 1,019,237 records, from 3 regions of the world, and contains 3 tasks. Based on the dataset, we trained NotaGPT, a music notation visual large language model. Specifically, we involve a pre-alignment training phase for cross-modal alignment between the musical notes depicted in music score images and their textual representation in ABC notation. Subsequent training phases focus on foundational music information extraction, followed by training on music score notation analysis. Experimental results demonstrate that our NotaGPT-7B achieves significant improvement on music understanding, showcasing the effectiveness of NOTA and the training pipeline.

1 Introduction

Music is expressed primarily in two forms: auditory music and symbolic music. Symbolic music can be represented in two-dimensional space through scores that display notes, rhythms, and dynamics, thereby guiding performers on how to play the music. It can also be expressed through lines of text sequences, effectively linearizing the complexity of music for ease of computer processing and programmatic manipulation. The evolution of Natural Language Processing (NLP) and multimodal interactions has provided valuable insights into the understanding and generation of music. With the advent of universal dialogue Multimodal Large Language Models (MLLMs) such as GPT-4(OpenAI, 2023), specialized models designed for various professional domains (Dey et al., 2024; Baez and Saggion, 2023), including music (e.g., MU-LLaMA (Liu et al., 2024)), have begun to proliferate. However, these works have only focused on the single modality of text, and in order to interact with multiple modalities, some MLLMs have been recently introduced. Nevertheless, these MLLM models mainly focus on the task of multimodal information extraction in the general domain, and rarely involve multimodal information extraction in the music domain, let alone the more advanced task of music notation understanding. Most existing datasets focus on specific symbols or audio (like ABC notation (Allwright, 2003), MIDI (Ryu et al., 2024), WAV (Sturm, 2013), and lyrics (Çano and Morisio, 2017)) and do not emphasize the visual modality, limiting their ability to enable MLLMs to comprehensively understand music notation, which is significantly important and has the largest data volume. Visual representations such as score images, which serve as a tangible record of music, provide an intuitive understanding and are crucial for professional music study. These images not only encapsulate the entirety of the score's information but also visually delineate its intricate structures.

To address the above limitations, we introduce NOTA, the first and largest comprehensive dataset designed to train and evaluate multimodal models in music notation understanding. Spanning three distinct global regions, NOTA encompasses over 1 million records of music scores. And it is structured around 3 pivotal tasks: music information extraction, cross-modal alignment test, and music notation analysis. These tasks cover various aspects of music, including music theory, composition, genres, musical ontological elements, and humanistic connotations. Our dataset is divided into two main parts: the training dataset and the test dataset. On the one hand, it provides train-

Train Dataset	
Alignment	28,125
Information Extraction	
T (Tune Title)	161,633
K (Key)	161,633
L (Unit Note Length)	161,633
M (Meter)	161,633
C (Composer)	161,633
ABC notation	161,636
Analysis	
Score Structure	150
Musical Style	300

Test Dataset	
Region	9,150
Information Extraction	
T (Tune Title)	1,851
K (Key)	1,851
L (Unit Note Length)	1,851
M (Meter)	1,851
C (Composer)	1,851
ABC notation	1,851
Analysis	
Score Structure	300
Musical Style	400

Figure 1: Data distribution of NOTA dataset.

ing materials for researchers in the community to train their own multimodal music models. On the other hand, it enables the evaluation of existing multimodal models' ability to understand music.

Based on this dataset, we trained a 7B model, NotaGPT, capable of understanding music notation across multiple modalities, including visual modalities. This training process comprises a pre-training phase focused on cross-modal alignment between the visual symbols in the music scores and their textual symbolic counterparts. This is followed by more specialized training phases that aim at foundational music information extraction, and music notation analysis.

Utilizing NOTA, we conducted comprehensive experiments on 17 mainstream multimodal large language models. Specifically, we input music score images and background information about the pieces, asking them to output basic information such as note lengths and key signatures or to perform analyses of the musical style and rhythm. Even the best-performing model, Gemini, achieved a music score information extraction rate of only 33.34%. In contrast, our 7B model, trained on our dataset, achieved 67.84%. The experimental results demonstrate the limitations in model performance caused by the lack of multimodal music datasets and highlight the effectiveness of our NOTA dataset and our training pipeline.

Our contribution can be summarized as follows: We introduced NOTA, the first and largest comprehensive multimodal music notation understanding dataset. This dataset encompasses 1,019,237 records from 3 distinct global regions and is dedicated to 3 tasks, addressing the resource limitation available for multimodal music notation understanding.

2 Related Work

2.1 Multimodal Benchmark

In the fields of NLP and multimodal interactions, traditional evaluation metrics predominantly focus on assessing specific capabilities of a model within singular task types(Goyal et al., 2017). For example, the GLUE (General Language Understanding Evaluation) (Sarlin et al., 2020) benchmark is a collection of diverse natural language understanding tasks designed to evaluate and advance the performance of models on a wide range of language comprehension challenges. These criteria either provide more dimensions of assessment (Guha et al., 2024; Sun et al., 2024)and advanced capabilities or employ sophisticated evaluation mechanisms (Wang et al., 2023; Valmeekam et al., 2024). For instance, the C-Eval (Huang et al., 2024b) benchmark addresses the gap in Chinese language data.

The evolution of evaluation benchmarks in NLP and multimodal fields has consequently influenced the benchmarks used in music evaluation. Presently, music evaluation metrics generally concentrate on distinct musical capabilities, such as music generation (Agostinelli et al., 2023; Melechovsky et al., 2023) and music information retrieval (Kong et al., 2020; Zhao and Guo, 2021). Some initiatives, such as ChatMusician (Yuan et al., 2024), attempt to unify tasks in music generation and comprehension, yet suffer from limited data volumes. Despite the rapid development of multimodal generative models, there is still a lack of data and benchmarks that can effectively evaluate the models' capabilities in understanding visual modality of music score images.

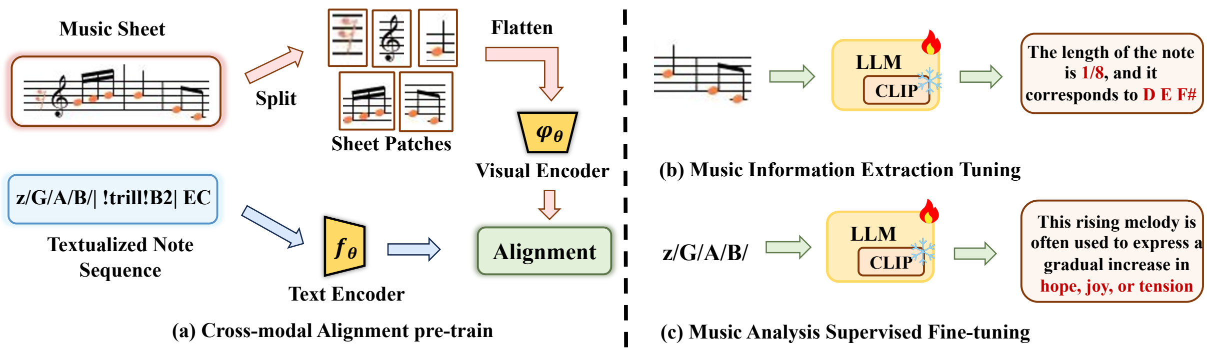

Figure 2: The figure shows the three-phase training process for NotaGPT-7B.

2.2 Generative Models for Music Understanding and Generation

With the advent of generative dialogue LLMs such as ChatGPT(OpenAI, 2022), alongside a series of universal dialogue MLLMs, specialized models designed for various professional domains (Li et al., 2024; Sun et al., 2022), including music (e.g., MU-LLaMA (Liu et al., 2024)), have begun to proliferate. As these MLLMs continue to evolve, music understanding capabilities have also been enhanced. For instance, current models like MusicAgent (Yu et al., 2023) and MusicLM (Agostinelli et al., 2023) have made remarkable progress in music comprehension and generation abilities.

Generative models for music understanding and generation can be broadly categorized into two modalities: audio music (Huang et al.; Copet et al., 2024) and symbolic music (Tian et al., 2024; Lu et al., 2023). The former predominantly incorporates audio modalities into large language models (Huang et al., 2024a) or employs diffusion models (e.g., JEN-1 (Li et al., 2023) and MeLoDy (Lam et al., 2024)) to process the audio components of music; the latter typically converts symbolic music information into sequences for integration into large language models (Yuan et al., 2024; Geerlings and Merono-Penuela, 2020).

The efficacy of these models hinges on precise instruction fine-tuning and cross-modal alignment (Geerlings and Merono-Penuela, 2020), utilizing specific musical datasets. Nevertheless, current generative music LLMs lack the ability to understand images of music scores in the visual modality.

2.3 Multimodal information extraction

Multimodal information extraction first searches for alignment in the two modalities connects them together, and then performs information extraction. It can be divided into two main categories: visual entity extraction and visual event extraction. In MORE (He et al., 2023), the objective is to predict relations between objects and entities based on both textual and image inputs. Visual event extraction can be further divided into situation recognition (Yatskar et al., 2016) and grounded situation recognition (Pratt et al., 2020). With the development of MLLMs, information extraction datasets for different tasks have also evolved (Wan et al., 2021; Yuan et al., 2023). However, there is still a lack of multimodal information extraction models and datasets specifically for the music domain.

3 NOTA Dataset

Our dataset is collected around three tasks: multimodal information extraction, multimodal alignment, and music notation analysis. We choose to use ABC notation to represent music scores. ABC notation encodes music into two parts: header and body. The first header is the reference number and the other headers are title T, time signature M, default note length L, key K, etc. The body mainly includes notes, bar lines, and so on.

Multimodal Music Information Extraction In this task, we collect a total of 1,185,761 data entries. Multimodal information extraction is divided into 6 subtasks: extracting ABC notation from corresponding images, and extracting specific information from the ABC notation, including T (tune title), K (key), L (unit note length), M (meter), and C (composer). We obtained 193,484 data entries from the ABC notation website, the vast majority of which are directly downloaded, and a small portion are scraped. After data cleaning, we only keep the ABC files that could generate the correct music score (we remove the original ABC file's comments, lyrics, and sequence numbers (X:)). We

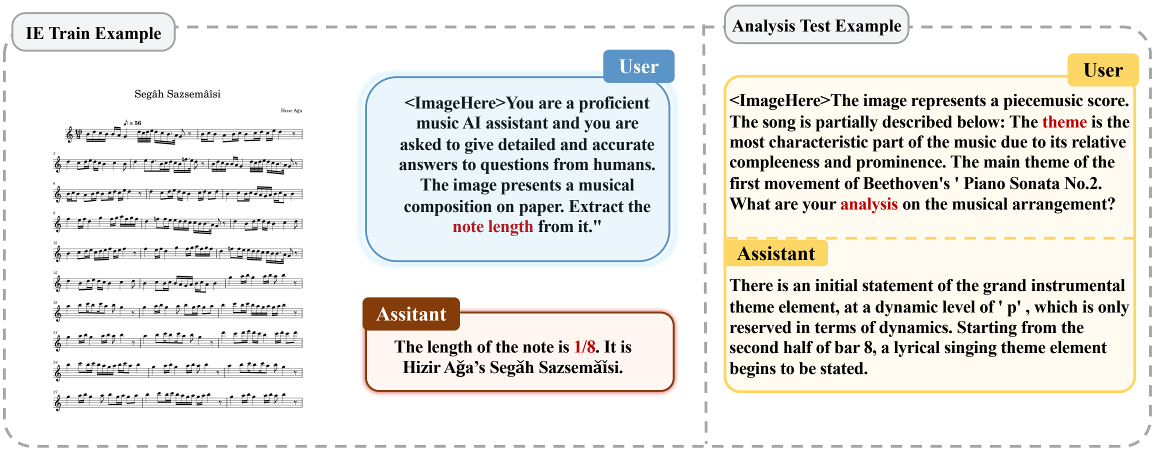

Figure 3: The left side of the figure shows an example of the information extraction task on the training dataset. The right side on the figure shows an example of music analysis for the test dataset.

then transform ABC files into MusicXML files and use MuseScore4 to generate music score images from the MusicXML files. Afterward, we divide each data entry into 6 data entries corresponding to 6 subtasks, resulting in 1,160,904 data entries.

In order to test whether MLLMs have a special tendency towards certain regions, we additionally collect nearly 4000*.krn* files from the internet, subsequently use the humdrum toolkit to convert them into ABC files, then filter and convert them into MusicXML files, generate music score from MusicXML files, and finally divide them into 6 extraction subtasks, obtaining a total of 24857 data entries with three regional labels: <China>, <Europe>, and <America>.

Each data sample includes the ABC notation information to extract, the corresponding music score images, the prompt used for extracting, and the gold answer. Data examples are in Figure 3.

Cross-modal Alignment In this task, we obtain 29,116 data entries. We highlight portions of the music score images, expecting that MLLMs can understand and extract the corresponding ABC notation content. Each music score image has 2 to 4 highlighted sections. For a music score image X_v and its associated content X_c, we sample a question X_q, which asks to extract the specific content of the image. With (X_v, X_c, X_q), we create a single-round instruction-following example:

$$\text{Human}: < ImageHere > \; X_q \; X_v \; < STOP >$$
$$\text{Assistant}: X_c \; < STOP > \quad (1)$$

Music Analysis This task includes analysis of score structure and musical styles. In terms of score structure analysis, it involves systematic analysis of various musical elements such as structure, melody, harmony, tonality, rhythm, tempo, dynamics, texture, etc. We integrate authoritative works on domestic and international music notation analysis. We obtain 250 questions on score structure analysis and 600 questions on musical style notation analysis. These questions cover the analysis of classic works from different countries (Germany, France, Italy, the UK, the United States, and so on) and different historical periods (from the Baroque period to the 20th century), involving various musical genres such as sonatas, symphonies, waltzes, and operas. Each data entry contains title, composer, the corresponding image, a description, and an analysis or structural breakdown.

Our dataset is divided into a train dataset and a test dataset. The train dataset has 998,976 samples, and the test dataset has 20,961 samples. More details are provided in Figure 1.

4 NotaGPT Training

We apply Mistral-7B (Jiang et al., 2023) as the base large language model and CLIP (Radford et al., 2021) as the vision encoder. Using the same network architecture as LLaVA (Liu et al., 2023a,b), the text model and the visual coder are connected through a linear projection layer. The model is first pre-trained with generalized domain multimodal datasets, which enables the model to understand images. Our music understanding training is mainly in three stages: phonogram-image notation-text alignment, music information extraction, and music comprehension, as shown in Figure 2.

Cross-modal Alignment At this stage, the primary goal is to achieve feature alignment between the musical notes depicted in music scores images and their textual representation in abc notation. Existing large vision models inherently lack this capability, as their pre-training does not include content specifically aligned with this requirement. Therefore, we have undertaken training modifications to enhance our model's performance. Specifically, we utilized the dataset introduced in section 3 to train the model. We have frozen the visual encoder and the language model components, focusing solely on training the two-layer MLP vision-language connector. This approach has enabled pre-alignment and endowed the model with the capability to recognize musical notes accurately.

Music Information Extraction Next, train the model to recognize the basic structure of music compositions and to extract relevant musical knowledge from images. Utilizing the training dataset described in section 3, we conducted fine-tuning of the entire model parameters while freezing the visual encoder component and training the remaining parts. Through this phase of training, the model's capability to extract musical information has significantly improved. It is now able to recognize fundamental elements of music scores such as beat types, note lengths, and key signatures from music score images.

Music Notation Analysis In the final phase, we further fine-tuned the model using supervised fine-tuning, thereby enhancing its capability to understand and generate music. This phase involved using the section 3 data to train the pre-trained projectors and the language model with full parameter adjustments. Post-training, the model has developed the ability to critically analyze music scores provided by users and perform complex tasks such as continuing a musical melody based on the preceding tune.

5 Experiments

5.1 Experiment Setup

Baselines We comprehensively assess 17 MLLMs, including API-based models and open-source models. The API-based models contain GPT-4V (GPT-4Vision-preview) (OpenAI, 2023), and Gemini model released by Google (Team et al., 2023). The open-source models contain LLaVA (Liu et al., 2023a,b) series, VisualGLM (Du et al., 2022), Qwen-VL (Bai et al., 2023) series, InternLM (Dong et al., 2024), InstructBLIP (Dai et al., 2024), and Yi-VL (Young et al., 2024) series.

Model	Levenshtein Distance
# Generative MLLM	
VisualGLM-6B	643.72
CogAgent-Chat	730.65
DeepSeek-VL-1.3B-Chat	316.85
DeepSeek-VL-7B-Chat	308.27
InstructBLIP-Vicuna-7B	355.60
Yi-VL-6B	561.47
Yi-VL-34B	522.07
LLaVA-v1.5-7B	667.08
LLaVA-v1.5-13B	147.47
LLaVA-v1.6-Vicuna-7B	807.75
LLaVA-v1.6-Vicuna-13B	918.94
LLaVA-v1.6-34B	770.58
Qwen-VL	439.82
Qwen-VL-Chat	625.16
#Generative MLLM with api-token	
Gemini-pro-vision	354.30
GPT-4V	655.45
# Our Models	
NotaGPT-7B	**59.47**

Table 1: Music cross-modal alignment evaluation.

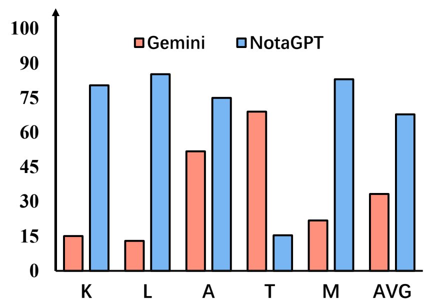

Figure 4: Extraction capabilities comparing between Gemini and NotaGPT-7B.

Training Details For pre-training, we utilized the alignment section 3 data conducting training 10 epoch with a learning rate of 2e-4. For supervised fine-tune training, we employed the train data in section 3, training 3 epochs with a learning rate of 2e-5 and a batch size of 32. All experiments are conducted on 8×80GB NVIDIA A100 SXM GPUs.

Evaluation Details The temperature parameter was set to 0 to ensure deterministic output. For each model, we performed 3 separate evaluations using the GPT-4 API. The final score is determined by averaging the results from these 3 assessments.

Model	Author	Title	K	L	M	Avg
CogAgent-Chat-hf	15.98	75.43	9.94	2.36	21.11	24.97
Cogvlm-Chat-hf	10.31	65.77	7.02	0.22	20.63	20.79
VisualGLM-6B	0.05	5.32	32.78	0.00	29.27	11.24
DeepSeek-VL-1.3B-Chat	15.98	0.11	4.75	0.00	22.84	8.74
DeepSeek-VL-7B-Chat	30.89	0.11	10.04	11.72	28.46	16.24
InstructBLIP-Vicuna-7B	0.43	5.67	7.67	0.00	1.84	3.12
Yi-VL-6B	46.27	17.82	10.37	9.13	5.02	17.72
Yi-VL-34B	60.85	0.22	13.55	14.36	11.18	20.03
LLaVA-v1.5-7B	54.81	25.16	11.56	11.50	28.54	26.31
LLaVA-v1.5-13B	6.86	34.23	4.7	0.59	28.22	14.92
LLaVA-v1.6-Vicuna-7B	38.88	59.56	6.97	1.94	23.95	26.26
LLaVA-v1.6-Vicuna-13B	11.99	60.69	7.99	0.92	7.84	17.89
LLaVA-v1.6-34B	15.66	62.31	11.18	1.46	28.22	23.76
MiniCPM-Llama3-V2_5	27.59	77.70	11.56	9.72	23.65	25.04
Qwen-VL	78.24	11.72	17.82	14.74	17.12	27.93
Qwen-VL-Chat	72.08	0.38	13.44	14.36	16.25	23.30
Gemini-pro-vision	51.83	69.03	15.08	13.02	21.87	33.34
GPT-4V	**82.24**	**77.95**	11.02	1.35	27.54	33.33
NotaGPT-7B	75.00	15.44	**80.45**	**85.26**	**83.08**	**67.84**

Table 2: Evaluation results of music information extraction task from the training dataset.

5.2 Evaluation Metrics

Closed-set tasks. ***(1)*** Such as *multimodal music information extraction*, performance is assessed using the weighted extraction rate. They are questions with definitive answers such as music titles and note lengths. Given a response sequence R and an answer sequence A across a dataset of n queries, the overall success of the extractions can be defined as:

$$Extraction\ \ Rate = \sum_{i=1}^{n} \delta\left([A_i \subseteq R_i], 1\right) \qquad (2)$$

where $\delta(x, y)$ is the Kronecker delta function, which equals 1 if $x = y$ and 0 otherwise. The condition $[A_i \subseteq R_i]$ evaluates to 1 if the answer sequence A_i is contained within the response sequence R_i, and 0 otherwise.

(2) Regarding the task of *converting images to abc notation text*, we utilize the Levenshtein Distance (Yujian and Bo, 2007) as evaluation metric. It refers to the minimum number of single-character operations required to transform model responses into answer sequence. Let D be a matrix of size $(|R| + 1) \times (|A| + 1)$, where $D[i][j]$ denotes the minimum edit distance between the first i characters of R and the first j characters of A. The subsequent values of D are computed using the recurrence relation:

$$D[i][j] = \min \begin{cases} D[i-1][j] + 1 & (delet) \\ D[i][j-1] + 1 & (insert) \\ D[i-1][j-1] + cost & (substitute) \end{cases} \qquad (3)$$

where cost is 0 if the characters $R[i-1]$ and $A[j-1]$ are the same, and 1 otherwise.

Open-set tasks. For *notation analysis* tasks with open-ended answers, we used 2 type assessment:

(1)Calculating using metrics. Our metrics are divided into two categories: semantic similarity and word matching. For semantic similarity, we use LSA, which measures the semantic similarity of text by computing the cosine similarity between vectors. For word matching, we use ROUGE-1, ROUGE-L, and METEOR, which respectively calculate the number of unigram matches, longest common subsequence matches, and synonym matches.

(2)Scoring using LLM as an evaluator. As existing studies (Zheng et al., 2023) demonstrated, strong LLMs can be good evaluators. We compare the analysis generated by NotaGPT-7B with the analysis generated by other models, and have GPT-4 (text model) evaluate the analysis from both models. The evaluation considers both the music itself and the music's background. The evaluation of the music itself includes aspects such as musical language (melody, tonality, rhythm, musical terminology, etc.), technique application, and composition style. The evaluation of the music's background includes considerations of the social, historical, and cultural context, including the composer's milieu, the background of the composition, and the ideology of the creation.

6 Results

Our experiment revolves around proving the effectiveness of NOTA in promoting music understand-

Model	LSA	ROUGE-1	ROUGE-L	METEOR	Avg
InternVL-Chat-v1.5	14.96	19.71	13.32	19.68	16.92
InternVL-14B-224px	3.28	5.30	4.63	4.18	4.35
VisualGLM-6B	10.36	21.61	13.21	18.19	15.84
DeepSeek-VL-7B-base	9.92	16.43	11.60	13.81	12.94
InstructBLIP-Flan-T5-xl	9.38	20.91	15.28	14.57	15.04
InstructBLIP-Flan-T5-xxl	7.64	17.55	12.32	14.96	13.12
InstructBLIP-Vicuna-7B	8.28	22.23	14.93	16.74	15.55
InstructBLIP-Vicuna-13B	8.37	20.29	14.18	14.17	14.25
MiniCPM-Llama3-V2_5	**16.26**	20.72	13.36	**20.83**	17.79
Yi-VL-6B	11.77	18.66	13.04	15.84	14.83
Yi-VL-34B	12.47	19.44	13.20	17.18	15.57
Qwen-VL	9.58	15.21	10.37	12.56	11.93
Qwen-VL-Chat	9.66	16.80	11.37	14.42	13.06
Gemini-pro-vision	15.88	22.21	15.09	20.31	**18.37**
GPT-4V	14.03	18.49	11.36	19.94	15.96
GPT4o	15.92	18.27	11.35	20.26	16.45
NotaGPT-7B	12.46	**22.63**	**15.53**	18.34	17.24

Table 3: Comparisons of analysis and form Evaluation (%). Part 1: Open-source models; Part 2: API-based models.

ing. In order to enable the model to ultimately achieve music understanding, we have broken down the experiment into three sub-experiments: multimodal information extraction, score image recognition, and music analysis. Multimodal information extraction only extracts the basic elements from the score image, such as author information, title, T, K, L, M and C. Score image recognition builds upon the basic element extraction, further extracting the music score in ABC notation form. Music analysis then, based on the extracted music score, conducts understanding and analysis, including score structure analysis and musical style analysis.

6.1 Music Information Extraction Evaluation

General comparison The evaluation results are presented in Table 2. We report the average extraction rate, with 23.53% of the models showing an effective precision lower than 10%. Additionally, 58.82% of the models have an accuracy approximately between 10% to 30% , and only 17.64% of the models achieve an accuracy exceeding 30%. Overall, NotaGPT-7B demonstrated the best performance among all the models evaluated, achieving an extracte rate of 67.84. These findings highlight the challenges of the NOTA test dataset.

Comparative analysis Figure 4 illustrates the comparative performance of NotaGPT-7B and Gemini in several subcategories of an information extraction task. NotaGPT-7B significantly outperforms Gemini in the tasks of Author, K, L, and M, demonstrating the effectiveness of the training data. NotaGPT-7B does not perform very well on the title extraction task, and after analyzing it, we found that it is because it mistakenly extracts author information as title information.

After training with the NOTA dataset, models of size 7B achieved substantial improvements in the categories K, L, and M, where performance was originally poor. These enhancements allowed them to surpass models of the same size and even those of larger sizes.

6.2 Music Cross-modal Alignment Evaluation

Table 1 presents the evaluation results. Overall, while high precision in music information extraction benefits cross-modal tasks, the relationship isn't simply linear. NotaGPT-7B consistently performs well, showcasing its strength in both extracting and aligning musical information. In contrast, while GPT-4V and Gemini-pro-vision score similarly in extraction tasks (around 33.34), they differ greatly in alignment accuracy, with Levenshtein distances of 655.45 and 354.30, respectively, suggesting that factors like model structure and optimization strategies also influence performance.

6.3 Music Score Analysis Evaluation

Metric evaluation Since the model's analysis and the standard answer cannot be completely identical, we evaluate the strength of the model's analysis capability of the recognized music score from the aspects of semantic similarity and word matching.

From the results in Table 3, in terms of the LSA metric, the performance of NotaGPT-7B is stronger

Model A	Type	Musical styles			Score Structures			C-Rate
		A win	Tie	B win	A win	Tie	B win	
InstructBLIP-Flan-T5-xxl	w/ Info.	5.00	33.50	61.50	1.34	33.56	65.10	96.56
	w/o Info.	5.50	39.00	55.50	1.34	26.84	71.81	96.27
InstructBLIP-Vicuna-7B	w/ Info.	1.00	25.00	74.00	2.68	32.89	64.43	98.28
	w/o Info.	1.50	36.00	62.50	2.01	28.86	69.12	98.28
InstructBLIP-Vicuna-13B	w/ Info.	1.00	26.50	72.50	1.34	30.87	67.79	98.85
	w/o Info.	2.00	35.00	63.00	0.13	23.48	75.17	98.28
InternVL-Chat-v1.5	w/ Info.	57.00	33.50	9.50	48.32	44.29	7.38	46.70
	w/o Info.	35.00	49.00	16.00	26.84	55.70	17.44	68.48
Qwen-VL	w/ Info.	24.50	45.00	30.50	16.11	39.60	44.30	79.08
	w/o Info.	0.50	33.50	66.00	0.67	19.46	79.87	99.43
VisualGLM-6B	w/ Info.	36.50	46.50	17.00	32.21	56.38	11.41	65.33
	w/o Info.	14.00	46.50	39.50	11.40	40.93	47.65	34.67
Yi-VL-6B	w/ Info.	36.00	40.50	23.50	30.20	49.66	20.14	66.47
	w/o Info.	94.00	3.50	2.50	13.42	38.92	47.65	40.40
GPT-4V	w/ Info.	69.50	25.00	5.50	55.70	33.56	10.74	36.39
	w/o Info.	52.00	34.00	14.00	32.88	49.66	17.44	56.16

Table 4: Results of models generating music analysis, evaluated by GPT-4 (text model). *Info.* means music background information, *A win* means in GPT-4's view, model A's response is better than model B's as evaluated by GPT-4; *tie* means the responses are equal; *B win* means model B's response is better. *C-Rate* means comparable rate between model B and model A.

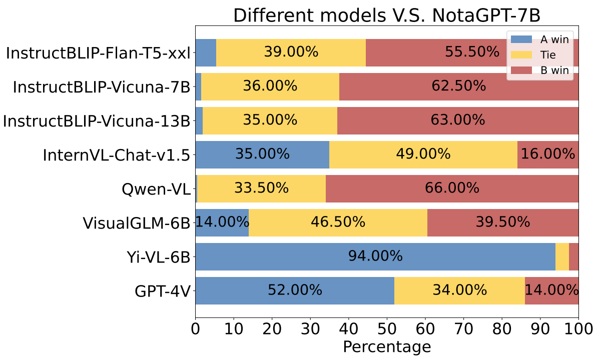

Figure 5: Visualization of evaluation results (w/o Info.) of all other models compared with our proposed NotaGPT model under GPT-4V.

than most models, including some models with larger parameter sizes than 7B, only second to a few open-source models with larger parameter sizes, as well as API-based models. In the metric of word matching , NotaGPT-7B achieves SOTA performance on 2/3 of the metrics.

NotaGPT-7B does not achieve the best performance on the LSA metric, on the one hand because the parameter size of NotaGPT-7B is only 7B, much smaller than the 25.5B of InternVL and the 34B of Yi-VL, which limits its capability; on the other hand, the base model of NotaGPT-7B does not use an instruction-tuned model like the Mistral-7B-Instruct series.

The results demonstrate the effectiveness of the NOTA test dataset, allowing the parameter-limited model NotaGPT-7B, after pre-training on the NOTA train dataset, to outperform models that have not been trained on the NOTA dataset in multimodal information extraction.

Analysis comparison Table 4 contains the comparison between analyses of different models, and all the model B are NotaGPT-7B. Based on the results, the appreciation generated by NotaGPT-7B is better or on par with 75% of the models. In comparison with most models, NotaGPT-7B's win rate is higher in the absence of music background information than with music background information. This performance can be attributed to NotaGPT-7B's training on a small set of music analysis data samples, which has endowed it with the capability to generally analyze musical scores and styles. It performs commendably even in prompts that lack background knowledge of the music piece.

7 Conclusion

In this study, we introduce NOTA, a large-scale music understanding dataset encompassing 3 tasks with over 1.1 million data entries. Based on the NOTA train dataset, we trained NotaGPT-7B, which demonstrates robust music notation understanding capability. We further assess 17 multimodal models' capabilities in music understanding. The results show the constraints that are caused by the lack of multimodal music datasets, emphasizing the significance of the NOTA dataset.

Limitations

Although NOTA makes substantial advancement in developing effective music understanding datasets, we are aware of typical limitations in MLLMs, including hallucinations and shallow reasoning. Our future efforts will focus on improving the fidelity and dependability of these models.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

James Allwright. 2003. Abc version of the nottingham music database. https://abc.sourceforge.net/NMD/index.html.

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Erion Çano and Maurizio Morisio. 2017. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence*, pages 118–124.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Schwartz. 2024. SOCIALITE-LLAMA: An instruction-tuned model for social scientific tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–468, St. Julian's, Malta. Association for Computational Linguistics.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Carina Geerlings and Albert Merono-Penuela. 2020. Interacting with gpt-2 to generate controlled and believable musical sequences in abc notation. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 49–53.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Liang He, Hongke Wang, Yongchang Cao, Zhen Wu, Jianbing Zhang, and Xinyu Dai. 2023. More: A multimodal object-entity relation extraction dataset with a benchmark evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4564–4573.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024a. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Max WY Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, et al. 2024. Efficient neural music generation. *Advances in Neural Information Processing Systems*, 36.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Peike Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. 2023. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. *arXiv preprint arXiv:2308.04729*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. 2023. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI Blog*.

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

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jesung Ryu, Seungyeon Rhyu, Hong-Gyu Yoon, Eunchong Kim, Ju Young Yang, and Taehyun Kim. 2024. Mid-fild: Midi dataset for fine-level dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 222–230.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Bob L Sturm. 2013. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. 2022. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jinhao Tian, Zuchao Li, Jiajia Li, and Ping Wang. 2024. N-gram unsupervised compoundation and feature injection for better symbolic music understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15364–15372.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, and Jeff Z Pan. 2021. Fl-msre: A few-shot learning based approach to multimodal social relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13916–13923.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Dingyao Yu, Kaitao Song, Peiling Lu, Tianyu He, Xu Tan, Wei Ye, Shikun Zhang, and Jiang Bian. 2023. MusicAgent: An AI agent for music understanding and generation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 246–255, Singapore. Association for Computational Linguistics.

Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11051–11059.

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

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095.

Yilun Zhao and Jia Guo. 2021. Musicoder: A universal music-acoustic encoder based on transformer. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27*, pages 417–429. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Appendix

A.1 Social Impact

The Nota-Eval dataset contains music from multiple regions and diverse cultural backgrounds. Not understanding the cultural context of the music may lead to misinterpretation of the music data, such as misreading the meaning and emotional expression of the music, as well as misjudging the characteristics and styles of the music.

A.2 Region-Level Evaluation

Table 5 presents the overall information extraction results for five information extraction tasks across 3 different regions using various models on our NOTA dataset. The experimental results indicate that the GPT-4V model significantly outperforms other models in music information extraction across different regions. For the five information extraction tasks in the regions of China and Europe, different models showed better performance compared to the America region. Additionally, there are noticeable differences in the information extraction capabilities of different models across the three regions. This suggests that different models have distinct preferences for understanding music from different regions, which may be related to the distribution of training data in these multimodal models.

A.3 Detailed Evaluation Metrics for Open-Set Tasks

Latent Semantic Analysis (LSA) is a technique in natural language processing and information retrieval that analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will appear in similar pieces of text. The core idea involves constructing a term-document

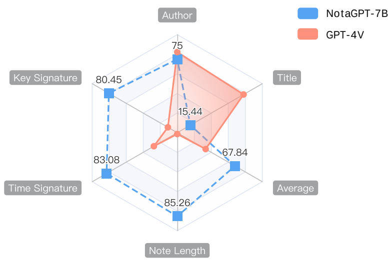

Figure 6: Comparing between GPT-4V and NotaGPT-7B.

Model	China	America	Europe	Avg
InternVL-14B-224px	0.00	0.00	0.15	0.05
InternVL-Chat-V1.5	0.48	5.56	1.81	2.61
VisualGLM-6B	8.66	2.53	10.36	6.64
DeepSeek-VL-1.3B-base	7.51	0.64	8.09	5.03
DeepSeek-VL-7B-base	4.08	0.32	1.94	2.31
InstructBLIP-Flan-T5-xl	0.46	0.17	2.46	0.69
InstructBLIP-Flan-T5-xxl	1.03	0.00	5.24	1.36
InstructBLIP-Vicuna-7B	3.57	0.47	5.89	2.80
InstructBLIP-Vicuna-13B	1.08	0.12	2.65	0.98
Yi-VL-6B	0.14	0.03	0.19	0.11
Yi-VL-34B	0.14	0.12	0.32	0.16
MiniCPM-Llama3-V2_5	6.79	5.97	11.39	7.26
Qwen-VL	2.35	1.31	1.88	1.88
Qwen-VL-Chat	0.26	0.47	0.13	0.32
GPT-4V	16.19	12.31	11.27	13.90

Table 5: Comparisons with SoTA for region-level Evaluation

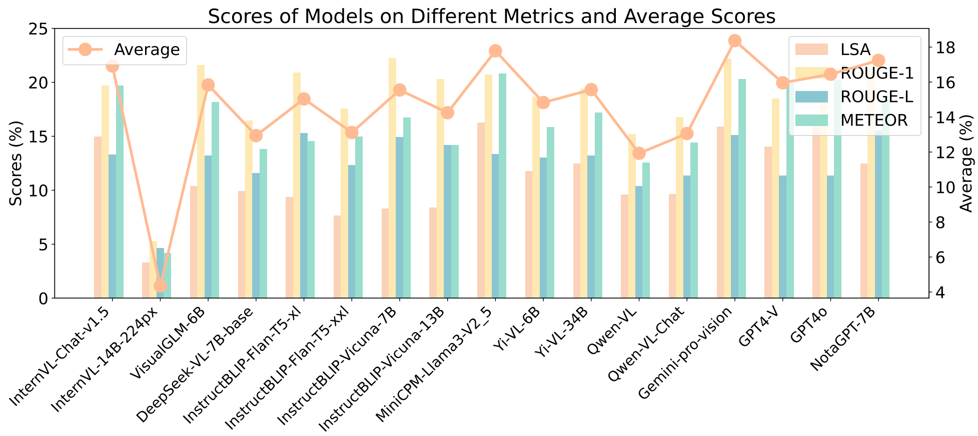

Figure 7: Music analysis figure.

matrix, which is then decomposed using singular value decomposition (SVD). The semantic similarity between texts is often measured using the cosine similarity between their vector representations. Let A be the term-document matrix, then LSA involves the following computation:

$$A \approx U_k \Sigma_k V_k^T$$

where:

- U_k represents the first k columns of U,
- Σ_k is the top $k \times k$ submatrix of Σ,
- V_k^T is the first k rows of V^T.

ROUGE-1 is a metric used to evaluate automatic summarization and machine translation software, focusing specifically on the overlap of unigrams (single words) between the system-generated summary or translation and a set of reference summaries. The ROUGE-1 score is calculated by counting the number of unigrams in the generated text that match the unigrams in the reference text and then normalizing this number by the total number of unigrams in the reference text, providing a measure of recall. ROUGE-N is a metric for evaluating text summarization and machine translation quality by measuring the overlap of N-grams between system-generated summaries and reference summaries. Specifically, ROUGE-1 is a variant of ROUGE-N where N equals 1, meaning it calculates the overlap using unigrams (individual words). ROUGE-1 focuses on assessing the recall of single words, providing a basic measure of content overlap and is widely used due to its simplicity and effectiveness in capturing essential content accuracy. ROUGE-N can be represented as:

$$\text{Rouge-N} = \frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

ROUGE-L measures the longest common subsequence (LCS) between a system-generated summary or translation and a set of reference texts. It is particularly useful for evaluating the fluency and the order of the text in summaries and translations. The LCS does not require consecutive matches but is a sequence where each word is in the same order in both texts. The score is computed by dividing the length of the LCS by the total length of the reference sequence, providing insights into the overall text structure retention.It can be represented as:

$$R_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{m}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X, Y)}{n}$$

$$F_{\text{lcs}} = \frac{(1 + \beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}}$$

METEOR, or the Metric for Evaluation of Translation with Explicit ORdering, is a metric for evaluating machine translation output by aligning it to one or more reference translations. Unlike other metrics, METEOR accounts for exact word matches, synonymy, and stemming. It calculates scores based on the harmonic mean of precision and recall, weighted towards recall. The inclusion of synonyms and stemming allows METEOR to perform a more nuanced assessment of language use than simple exact matching. The METEOR score is calculated as follows:

$$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty})$$

where:

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

$$\text{Penalty} = 0.5 \times \left(\frac{\text{number of chunks}}{\text{number of unigrams in candidate translation}}\right)^3$$

In these equations:

- P is the precision,
- R is the recall,
- Chunks are contiguous sequences of words that are in the same order in both the candidate and the reference but are possibly interspersed with non-matching words.