LP-MusicCaps: LLM-BASED PSEUDO MUSIC CAPTIONING

SeungHeon Doh[♭] **Keunwoo Choi**[♮] **Jongpil Lee**[♯] **Juhan Nam**[♭]
[♭] Graduate School of Culture Technology, KAIST, South Korea
[♮] Gaudio Lab, Inc., South Korea
[♯] Neutune, South Korea
{seungheondoh, juhan.nam}@kaist.ac.kr, keunwoo@gaudiolab.com, jongpillee@neutune.com

ABSTRACT

Automatic music captioning, which generates natural language descriptions for given music tracks, holds significant potential for enhancing the understanding and organization of large volumes of musical data. Despite its importance, researchers face challenges due to the costly and time-consuming collection process of existing music-language datasets, which are limited in size. To address this data scarcity issue, we propose the use of large language models (LLMs) to artificially generate the description sentences from large-scale tag datasets. This results in approximately 2.2M captions paired with 0.5M audio clips. We term it **L**arge Language Model based **P**seudo music caption dataset, shortly, **LP-MusicCaps**. We conduct a systemic evaluation of the large-scale music captioning dataset with various quantitative evaluation metrics used in the field of natural language processing as well as human evaluation. In addition, we trained a transformer-based music captioning model with the dataset and evaluated it under zero-shot and transfer-learning settings. The results demonstrate that our proposed approach outperforms the supervised baseline model. [1]

1. INTRODUCTION

Music captioning is a music information retrieval (MIR) task of generating natural language descriptions of given music tracks. The text descriptions are usually sentences, distinguishing the task from other music semantic understanding tasks such as music tagging. Recently, there have been some progress in music captioning including track-level captioning [1, 2] and playlist-level captioning [3–6]. These approaches usually utilize a deep encoder-decoder framework which is originally developed for neural machine translation [7]. Choi *et al.* [3] used a pre-trained music tagging model as a music encoder and an RNN layer initialized with pre-trained word embeddings for text generation. Manco *et al.* [1] introduced a temporal attention mechanism for alignment between audio and text by pairing a pre-trained harmonic CNN encoder [8] with an LSTM layer. Gabbolini *et al.* [5] generated playlist titles and descriptions using pre-trained GPT-2 [9].

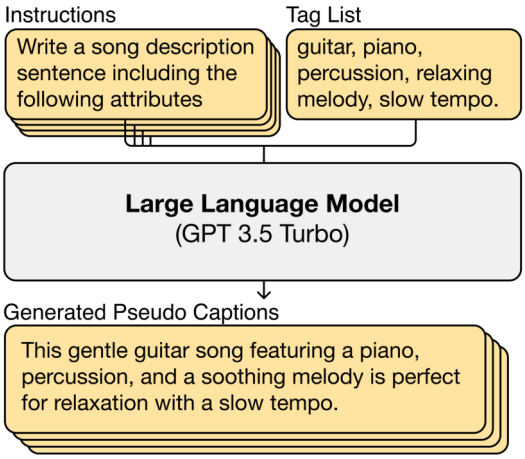

Figure 1. The generation process of pseudo captions by feeding a large language model with instructions and manually-annotated labels.

Currently, the primary challenge of track-level music captioning is the scarcity of large-scale public datasets. Manco *et al.* [1] used private production music datasets. Huang *et al.* [10] also used a private dataset with 44M music-text pairs on YouTube, but this approach is hardly reproducible or affordable for other researchers. To address this data issue, a community-driven data collection initiative has been proposed [11]. As of now, the only publicly available dataset for track-level music captioning is MusicCaps [12], which includes high-quality music descriptions from ten musicians. However, it is limited to 5521 music-caption pairs as it was originally created as an evaluation set for a text-prompt music generator.

With the scale of the aforementioned datasets, it remains difficult to train a music captioning model successfully. A workaround for this situation is to use music tagging datasets and generate sentences with tag concatenation [2, 13] or prompt template [14]. As relying on tagging datasets, however, the tag-to-sentence approaches would have the same limitation tagging datasets have. For example, high false-negative rates of tagging datasets [15]. Tag-

[1] Our dataset and codes are available at https://github.com/seungheondoh/lp-music-caps

ging datasets also has some typical issues text data have, for example, synonyms, punctuation, and singular/plural inconsistencies. Without proper treatment, these can limit the performance of the corresponding music captioning models.

A potential solution is to use strong language models, i.e., large language models (LLMs). LLMs refer to the recent large-scale models with over a billion parameters that exhibit strong few-shot and zero-shot performance [9, 16]. Large language models are usually trained with text data from various domains such as Wikipedia, GitHub, chat logs, medical articles, law articles, books, and crawled web pages [17]. When successfully trained, they demonstrate an understanding of words in various domains [9]. There have been similar and successful use cases of LLMs for general audio understanding [18] and music generation [19].

Motivated by the recent success of LLMs, we propose creating a music captioning dataset by applying LLMs carefully to tagging datasets. Our goal is to obtain captions that are i) semantically consistent with the provided tags, ii) grammatically correct, and iii) with clean and enriched vocabulary. This dataset-level approach is rather pragmatic than sophisticated; it alleviates the difficulty of music captioning tasks not by theory or model, but by data. The aforementioned ambiguous aspects of the music captioning task are addressed by the powerful LLMs that cost reasonably [20], considering the training cost music researchers would spend otherwise. Once the creation is complete, it is straightforward to train some music captioning models by supervised learning.

There are some existing works in the pseudo-labeling using language models. Huang *et al.* [19] introduced the MuLaMCap dataset, which consists of 400k music-caption pairs generated using the large language model and the music-language joint embedding model. They utilized a large language model (LaMDA [21]) to generate 4M sentences using 150k song metadata as input in the format of `{title} by {artist}`. Then the text and music-audio joint embedding model, MuLan, calculates the similarity between music and generated captions, annotating pairs with high similarity [10]. However, it is not possible to reproduce or evaluate this work as the adopted language model as well as the final music-audio embedding model are not publicly available. Moreover, using metadata has some issues – a popularity-biased, limited coverage and a low reliability – as we discuss later in Section 2.1. Wu *et al.* [22] introduce keyword-to-caption augmentation (K2C Aug) to generate captions based on the ground truth tags of audio clips in AudioSet. They used a pre-trained T5 model without any instruction. Finally, Mel *et al.*. [18] introduce WavCaps, a 400k audio captioning dataset using ChatGPT [23]. However, previous approaches only reported task performance and did not directly evaluate the quality of generated captions.

We propose a solution in this paper with three-fold key contribution. First, we propose an LLM-based approach to generate a music captioning dataset, **LP-MusicCaps**. Second, we propose a systemic evaluation scheme for music captions generated by LLMs. Third, we demonstrate that models trained on LP-MusicCaps perform well in both zero-shot and transfer learning scenarios, justifying the use of LLM-based pseudo-music captions.

2. PSEUDO CAPTION GENERATION USING LARGE LANGUAGE MODELS

In this section, we introduce how music-specific pseudo captions are created using a large language model in the proposed method.

2.1 Large Language Model for Data Generation

We first take multi-label tags from existing music tagging datasets. The list of tags are appended with a carefully written task instruction as an input (prompt) to a large language model. The model then generates and returns sentences that (may) describe the music in a way the task instruction conditions. Table 1 shows examples of generated captions according to multi-label tags and task instructions. For the language model, we choose GPT-3.5 Turbo [23] for its strong performance in various tasks. During its training, it was first trained with a large corpus and immense computing power, then fine-tuned by reinforcement learning with human feedback (RLHF) [24] for better interaction with given instruction. As a result, GPT-3.5 Turbo demonstrates state-of-the-art zero-shot abilities in understanding, reasoning, and generating human-like responses to natural language inputs.

Since LLMs contain a wide range of information, music captions may be generated based on some famous musical entities such as the artist name or album name. However, LLMs may generate inaccurate text in a confident tone which is hard to detect without ground truth. This issue, known as hallucination, can be a fun aspect when using LLMs for creative purposes [25]. However, hallucination should be avoided in an application like ours as the resulting captions should be factual. Therefore, we do not use any metadata unlike a previous work [19]. We also added a question to measure hallucination in the proposed evaluation scheme.

2.2 Task Instruction Design

Our proposed caption generation follows the formulation: $\tilde{y}_{\text{cap}} = f_{\text{LLM}}(y_{\text{tag}}, i)$, where y_{tag} and $\tilde{y}_{\text{cap}}$ refer to the multi-label tag and the generated caption, respectively, and i is the task instruction provided. Given that the output can vary based on the task instruction, even with the same model and input, task instructions become a crucial aspect of data generation. Therefore, we define four different tasks and generate captions accordingly.

Writing: *Write a song description sentence including the following attributes.* `{input tags}`

Summary: *Write a single sentence that summarizes a song with the following attributes. Don't write the artist name or album name.* `{input tags}`

Input tags video game theme, no singer, instrumental, analog sounding, small keyboard, beatboxing, playful, cheerful, groovy
Result of "Writing" Instruction This instrumental track has a joyful and playful vibe, perfect for a video game theme. With no singer, the analog-sounding music features a small keyboard and beatboxing, creating a groovy and cheerful atmosphere.
Result of "Summary" Instruction A cheerful and groovy instrumental track with playful beatboxing and an analog sound, featuring a small keyboard and video game-inspired theme.
Result of "Paraphrase" Instruction This upbeat instrumental tune captures the essence of a fun-filled video game theme, incorporating analog-sounding beats, playful beatboxing rhythms, and a groovy small keyboard melody, all without the need for a singer.
Result of "Attribute Prediction" Instruction {"new_attribute": ["8-bit sound", "chiptune style", "retro vibe"], "description": "This instrumental tune is straight out of a video game with its analog sounding melodies and small keyboard tinkles. Beatboxing adds a playful element to the groovy, cheerful vibe. Reminiscent of classic 8-bit sound and chiptune style, this retro vibe is sure to put a smile on your face."}
Ground Truth This is a video game theme cover. The theme belongs to the Super Mario franchise. The main theme is being played on an analog sounding small keyboard. There is an added rhythmic background of beatboxing in this version. The atmosphere is playful. This piece could be used in the background of arcade gaming social media content.

Table 1. An example of generated captions from MusicCaps dataset.

Paraphrase: *Write a song description sentence including the following attributes. Creative paraphrasing is acceptable.* `{input tags}`

Attribute Prediction: *Write the answer as a Python dictionary with new_attribute and description as keys. For new_attribute, write new attributes that show high co-occurrence with the following attributes. For description, write a song description sentence including the following attributes and new attributes.* `{input tags}`

In every instruction, we add 'include / with the following attributes' to prevent hallucination. The "Writing" task instruction is a simple prompt that uses tags to generate a sentence. The "Summary" task instruction aims to compress information into a short length. The "Paraphrase" task instruction expands the vocabulary. Finally, the "Attribute Prediction" task instruction predicts new tags based on tag co-occurrence in large corpora (i.e. the training data of GPT-3.5 Turbo), which is expected to address the issue of high false-negative rates in existing tagging datasets while mitigating the risk of hallucination. In this instruction, 'new attributes' exists to bridge the description and the input, and we only use the 'description' as caption.

3. EVALUATION OF PSEUDO CAPTIONS

It is crucial to ensure the quality of generated captions, especially since they are supposed to be used as ground truth. In this section, we introduce a holistic evaluation scheme that includes objective and subjective assessment – and its result on the captions from the proposed method.

3.1 Objective Evaluation

We conduct evaluation on the generated captions using MusicCaps dataset [12]. It has audio (x), tag list (y_{tag}), and ground truth caption (y_{cap}). The pseudo captions ($\tilde{y}_{\text{cap}}$) are generated with four pre-defined instructions as explained in Section 2.2 for all items in the evaluation split. During the evaluation, the generated captions are compared to the ground truth captions with respect to *n*-gram, neural metrics. We also report diversity metrics.

Following the previous work [5], we measure four *n*-gram metrics [26–28]: BLEU1 to 4 (B1, B2, B3, B4), METEOR (M), and ROUGE-L (R-L). They are all based on *n*-gram precision and recall between the ground truth and generated captions. These metrics capture different aspects of the caption quality. BLEU and METEOR focus on *n*-gram overlap between the generated and ground truth captions, while ROUGE-L measures the longest common subsequence between the two.

In addition, we use BERT-Score (BERT-S) based on pre-trained BERT embeddings to represent and match the tokens in the ground truth with respect to the generated caption [29]. By computing the similarity between the BERT embeddings of each token, BERT-Score can better capture the semantic similarity between the generated and ground truth captions than *n*-gram metrics; as it is more robust to synonyms, paraphrasing, and word order variations.

Finally, we evaluate the diversity of the generated captions by measuring how many different words are used. $novel_v$ indicates the percentage of new vocabulary in generated captions that are not among the training vocabulary. *Vocab* is the number of unique words used in all the generated captions. It is worth noting that diversity metrics are generally considered as subsidiaries and do not capture the overall quality of the generated captions.

3.2 Subjective Evaluation

Following the previous work [12], we set up an A-vs-B human rating task, in which a participant is presented with a 10-second single music clip and two text descriptions. We randomly selected 240 music samples from the MusicCaps evaluation dataset. Since the research goal is to generate

Methods	LM	Params	Supervised Metrics							Diversity Metrics		Length
			B1↑	B2↑	B3↑	B4↑	M↑	R-L↑	BERT-S↑	Vocab↑	Novel_{v}↑	Avg.Token
Baseline												
Tag Concat [2, 13]	-	-	20.25	13.57	8.64	5.42	23.24	19.52	86.24	3506	46.92	20.6±11.2
Template [14]	-	-	25.41	16.15	10.00	6.15	25.57	21.36	87.92	3507	46.93	25.6±11.2
K2C Aug. [22]	T5	220M	6.07	3.01	1.58	0.85	14.23	17.92	86.33	3760	**67.66**	14.7±5.1
Proposed Instruction												
Writing	GPT3.5	175B+	**36.84**	**19.85**	**11.37**	**6.74**	31.44	25.36	89.26	5521	56.17	44.4±17.3
Summary	GPT3.5	175B+	26.12	14.58	8.80	5.52	27.58	**25.83**	**89.88**	4198	49.52	28.6±10.7
Pharapase	GPT3.5	175B+	36.51	18.73	10.33	5.87	30.36	23.40	88.71	6165	59.95	47.9±18.7
Attribute Prediction	GPT3.5	175B+	35.26	18.16	9.69	5.41	**34.09**	23.19	88.56	**6995**	63.16	66.2±21.6

Table 2. Performance of existing pseudo caption generation methods and the proposed method. LM stand for the language model. Avg.Token stand for the average number of token per caption.

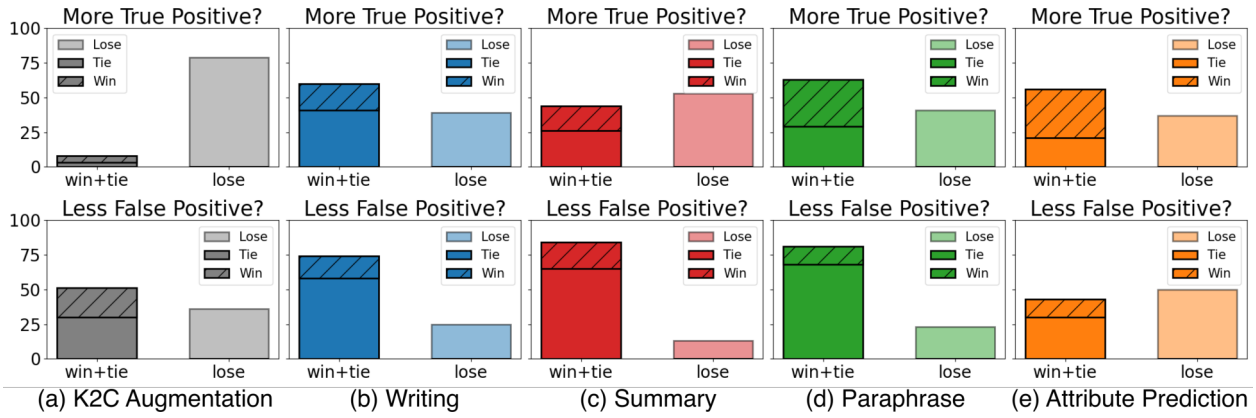

Figure 2. A-vs-B test results. Each method is compared to ground truth in terms of having more true positives and fewer false positives. The proposed methods (b, c, d, e) show comparable **win+tie** performance to ground truth.

music captions that can be used as pseudo-ground truth, one description is always fixed to the ground truth and the other is chosen from 5 types of generated captions including the K2C Augmentation [22] and the four proposed instruction methods. This yields up to 1200 (= 240 x 5) questions. We hired 24 participants who are music researchers or professionals in the music industry. Each of them rated 20 randomly selected questions. As a result, we collected a total of 480 ratings. The rater was asked to evaluate caption quality on two different aspects: (Q1) *More True Positive*: which caption describes the music with more accurate attributes? (Q2) *Less False Positive*: which caption describes the music less wrong? For example, if a method produces long and diverse sentences with many music attributes, it may be advantageous for Q1 but disadvantageous for Q2. Conversely, if a method conservatively produces short sentences with few music attributes, it may be advantageous for Q2 but disadvantageous for Q1. We determine the ranking of conditions by counting the number of wins, ties, and loses in the pairwise tests.

3.3 Results

We compare our LLM-based caption generation with two template-based methods (tag concatenation, prompt template [2]) and K2C augmentation [22]. In Table 2, we present the captioning result for MusicCaps [12] evaluation set. When comparing our proposed method with existing methods, we observe significant differences in *n*-gram metrics. This is because the tag concatenation fails to complete the sentence structure. In the case of K2C Augmentation, due to the absence of instruction, the input tag is excluded from the generated caption, or a sentence unrelated to the song description sentence is created. In contrast, the template-based model shows improved performance as the musical context exists in the template. We next consider diversity metric with BERT-Score. Our proposed method shows higher values in BERT-Score while generating diverse vocabularies. This indicates that the newly created vocabulary does not harm the music semantics.

Comparing within the proposed different task instructions, we can observe that each instruction performs a different role. "Writing" shows a high n-gram performance as it faithfully uses input tags to generate captions. "Summary" has the smallest average number of tokens due to its compression of information, but it shows competitive performance in ROUGE-L which is specialized to summarizing, as well as the highest BERT-Score. "Paraphrase" generates many synonyms, resulting in a large vocabulary size and the use of novel vocabulary. "Attribute Prediction" predicts new tags based on the co-occurrence of tags. This instruction shows lower performance in BLEU but competitive results in METEOR, which utilizes a thesaurus, such as WordNet, to consider the accuracy scores of words with similar meanings, indicating that newly predicted tags have similar semantic with ground truth.

Figure 2 shows the subjective A-vs-B test results. Each

[2] Template example: the music is characterized by `{input tags}`

Dataset	# item	Duration (h)	C/A	Avg. Token
General Audio Domain				
AudioCaps [30]	51k	144.9	1	9.0±N/A
LAION-Audio [22]	630k	4325.4	1-2	N/A
WavCaps [18]	403k	7568.9	1	7.8±N/A
Music Domain				
MusicCaps [12]	6k	15.3	1	48.9±17.3
MuLaMCap* [19]	393k	1091.0	12	N/A
LP-MusicCaps-MC	6k	15.3	4	44.9±21.3
LP-MusicCaps-MTT	22k	180.3	4	24.8±13.6
LP-MusicCaps-MSD	514k	4283.1	4	37.3±26.8

Table 3. Comparison of audio-caption pair datasets. C/A stands for the number of caption per audio. *Although we include MuLaMCap in the table for comparison, it is not publicly accessible.

method is compared to the ground truth in terms of having more true positives (Q1) and fewer false positives (Q2). For the first question, compared to the baseline K2C augmentation, the proposed methods using the instructions show an overwhelmingly higher *win+tie* score. This indicates the importance of music-specific instructions when utilizing LLM. In particular, "Paraphrase" and "Attribute Prediction" achieve high *win* scores by incorporating new information that is different from the existing vocabulary. In the second question, all caption generation methods except "Attribute Prediction" show higher *win+tie* scores than *lose* scores. This advocates the trustworthiness of LLM-based caption generation as it shows a similar or less false-positive rate to the ground truth. With its longest average length, "Attribute Prediction" turns out to be 'too creative' and shows a slightly higher false-positive rate than the ground truth.

4. DATASET: LP-MusicCaps

Based on the proposed pseudo caption generation method, we introduce LP-MusicCaps, an LLM-based Pseudo music caption dataset. We construct the music-to-caption pairs using three existing multi-label tag datasets and four task instructions. The data sources are MusicCaps [12], Magnatagtune [31], and Million Song Dataset [32] ECALS subset [13]. We respectively refer to them as MC, MTT, and MSD. MC contains 5,521 music examples, [3] each of which is labeled with 13,219 unique aspects written by music experts. MTT [31] consists of 26k music clips from 5,223 unique songs including genre, instrument, vocal, mood, perceptual tempo, origin, and sonority features. We used the full 188 tag vocabulary and did not generate captions for tracks that do not have associated tags (decreased to 22k). MSD consists of 0.52 million 30-second clips and 1054 tag vocabulary [13]. The tag vocabulary covers various categories including genre, style, instrument, vocal, mood, theme, and culture. Each dataset uses an average of 10.7 / 3.3 / 10.2 labels per music clip for generating pseudo captions, respectively.

Table 3 provides a comparison of statistics between the LP-MusicCaps family and other audio-caption pair datasets. When comparing the two domains, AudioCaps [30] and MusicCaps have high-quality human annotated captions, but they have fewer captions with shorter audio duration. When comparing large-scale datasets, the music domain lacks available datasets compared to the general audio domain (such as LAION-Audio [22] and WavCaps [18]). Although MuLaMCap has an overwhelming amount of annotated captions, it is not publicly available. In contrast, LM-MusicCaps is publicly accessible and provided with various scales. LP-MusicCaps-MC has a similar caption length to manually written captions while having four times more captions per audio. LP-MusicCaps-MTT is a medium-sized dataset with audio download link, and LP-MusicCaps-MSD has the largest audio duration among various captions in the music domain caption dataset.

[3] We only use 5495 out of the total due to the loss of 26 data samples.

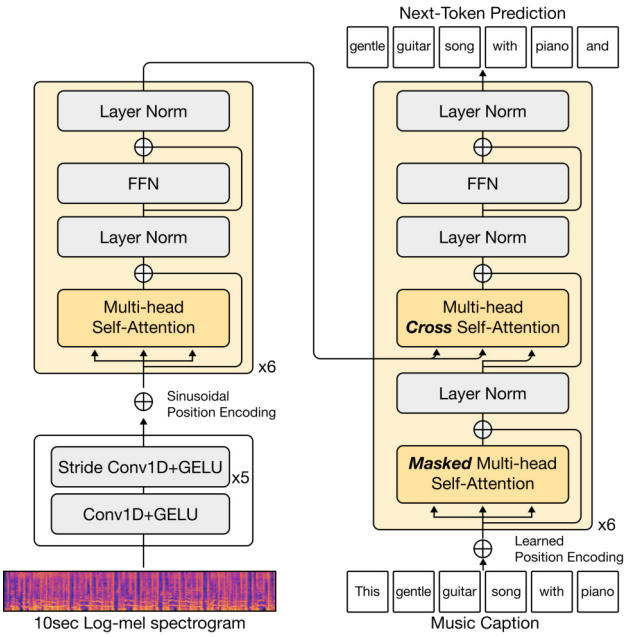

Figure 3. A cross-modal encoder-decoder architecture.

5. AUTOMATIC MUSIC CAPTIONING

We trained a music captioning model and evaluated it under zero-shot and transfer-learning settings. This section reports the experimental results.

5.1 Encoder-Decoder Model

We used a cross-modal encoder-decoder transformer architecture that has achieved outstanding results on various natural language processing tasks [33], lyrics interpretation [34], and speech recognition [35], as shown in Figure 3. Similar to Whisper [35], the encoder takes a log-mel spectrogram with six convolution layers with a filter width of 3 and the GELU [36] activation function. With the exception of the first layer, each convolution layer has a stride of two. The output of the convolution layers is combined with the sinusoidal position encoding and then processed by the encoder transformer blocks. Following the $\text{BART}_{\text{base}}$ architecture, our encoder and decoder both have 768 widths and 6 transformer blocks. The decoder

	Supervised Metrics							Diversity Metrics			Length
Model	B1↑	B2↑	B3↑	B4↑	M↑	R-L↑	BERT-S↑	Vocab↑	Novel_v↑	Novel_c↑	Avg.Token
Baseline											
Supervised Model	28.51	13.76	7.59	4.79	20.62	19.22	87.05	2240	0.54	69.00	46.7±16.5
Zeroshot Captioning											
Tag Concat [2, 13]	4.33	0.84	0.26	0.00	3.10	2.01	79.30	802	46.38	100.00	23.8±12.1
Template [14]	7.22	1.58	0.46	0.00	5.28	6.81	81.69	787	45.24	100.00	25.8±12.4
K2C-Aug [22]	7.67	2.10	0.49	0.10	7.94	11.37	82.99	**2718**	**81.97**	100.00	19.9±7.6
LP-MusicCaps (**Ours**)	**19.77**	**6.70**	**2.17**	**0.79**	**12.88**	**13.03**	**84.51**	1686	47.21	100.00	45.3±28.0
Tansfer Learning											
Tag Concat [2, 13]	28.65	14.68	8.68	5.82	21.88	21.31	87.67	1637	3.30	96.07	41.8±14.3
Template [14]	28.41	14.49	8.59	5.78	21.88	21.25	87.72	1545	**3.62**	**96.77**	41.1±13.2
K2C-Aug [22]	**29.50**	**14.99**	8.70	5.73	21.97	20.92	87.50	**2259**	1.42	84.95	44.1±15.0
LP-MusicCaps (**Ours**)	29.09	14.87	**8.93**	**6.05**	**22.39**	**21.49**	**87.78**	1695	1.47	96.06	42.5±14.3

Table 4. Music captioning results on the MusicCaps eval-set. Avg.Token stands for the average number of token per caption.

processes tokenized text captions using transformer blocks with a multi-head attention module that includes a mask to hide future tokens for causality. The music and caption representations are fed into the cross-modal attention layer, and the head of the language model in the decoder predicts the next token autoregressively using the cross-entropy loss, formulated as: $\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{1:t-1}, x)$ where x is the paired audio clip and y_t is the ground truth token at time t in a caption with length T.

5.2 Experimental Setup

To evaluate the impact of the proposed dataset on the music captioning task, we compare a supervised model trained on the MusicCaps [12] training split and a pre-trained model trained on an LP-MusicCaps-MSD dataset. For the pre-trained model, we perform both a zero-shot captioning task that does not use any MusicCaps [12] dataset and a fine-tuning task that updates the model using MusicCaps [12] training split. For comparison with other pseudo caption generation methods, we report results on baseline models trained with the same architecture and amount of audio, but different pseudo captions. In addition to all the metrics we used in Section 3.1, we compute *Novel*$_c$, the percentage of generated captions that were not present in the training set [37]. It measures whether the captioning model is simply copying the training data or not.

For all the experiments, the input of the encoder is a 10-second audio signal at 16 kHz sampling rate. It is converted to a log-scaled mel spectrogram with 128 mel bins, 1024-point FFT with a hann window, and a hop size of 10 ms. All models are optimized using AdamW with a learning rate of 1e-4. We use a cosine learning rate decay to zero after a warmup over the first 1000 updates. For the pre-training dataset, we use 256 batch-size and the models are trained for 32,768 updates. We adopt a balanced sampling [38], which uniformly samples an anchor tag first and then selects an annotated item. For supervised and transfer learning, we use a 64 batch size, 100 epochs. We use beam search with 5 beams for the inference of all models.

5.3 Results

When comparing within zero-shot captioning models, the model trained on the proposed LP-MusicCaps dataset shows a strong performance in general. The model using tag concatenation shows the lowest performance as it fails to generate musical sentences. In case of the model using a prompt template, it demonstrates a slightly higher BERT-Score, while still exhibiting poor performance in terms of n-gram metrics due to its limited vocabulary. The model using K2C augmentation outperforms the other two methods but still falls short due to its lack of a musical context. In general, zero-shot models does not perform as well as the supervised baseline in most of the metrics with few exceptions.

Among the transfer captioning models, the model with LP-MusicCaps pre-training achieves strong performance overall by winning in the BERT-Score and most of the n-gram metrics. It is noteworthy that our proposed model shows a meaningful increase in BERT-Score compared to the supervised model. This improvement is likely a result of successful semantic understanding rather than word-to-word matching. Moreover, by the improvement of Novel_c, the LP-MusicCaps model demonstrates that it can generate new captions instead of repeating the phrases in the training dataset. This advantage is observed in both the zero-shot and supervised tasks in transfer learning models.

6. CONCLUSION

We proposed a tag-to-pseudo caption generation approach with large language models to address the data scarcity issue in automatic music captioning. We conducted a systemic evaluation of the LLM-based augmentation, resulting in the creation of the LP-MusicCaps dataset, a large-scale pseudo-music caption dataset. We also trained a music captioning model with LP-MusicCaps and showed improved generalization. Our proposed approach has the potential to significantly reduce the cost and time required for music-language dataset collection and facilitate further research in the field of connecting music and language, including representation learning, captioning, and generation. However, further collaboration with the community and human evaluation is essential to enhance the quality and accuracy of the generated captions. Additionally, we believe that exploring the use of LLMs for other topics under music information retrieval and music recommendation could lead to novel and exciting applications.

7. REFERENCES

[1] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Muscaps: Generating captions for music audio," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

[2] T. Cai, M. I. Mandel, and D. He, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020.

[3] K. Choi, G. Fazekas, B. McFee, K. Cho, and M. Sandler, "Towards music captioning: Generating music playlist descriptions," in *International Society for Music Information Retrieval Conference (ISMIR), Late-Breaking/Demo*, 2016.

[4] S. Doh, J. Lee, and J. Nam, "Music playlist title generation: A machine-translation approach," in *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MuSA)*, 2021.

[5] G. Gabbolini, R. Hennequin, and E. Epure, "Data-efficient playlist captioning with musical and linguistic knowledge," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[6] H. Kim, S. Doh, J. Lee, and J. Nam, "Music playlist title generation using artist information," in *Proceedings of the AAAI-23 Workshop on Creative AI Across Modalities*, 2023.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[8] M. Won, S. Chun, O. Nieto, and X. Serrc, "Data-driven harmonic filters for audio representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Proceedings of the Advances in neural information processing systems (NeurIPS)*, 2020.

[10] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "MuLan: A joint embedding of music audio and natural language," in *International Conference on Music Information Retrieval (ISMIR)*, 2022.

[11] I. Manco, B. Weck, P. Tovstogan, M. Won, and D. Bogdanov, "Song describer: a platform for collecting textual descriptions of music recordings," in *International Conference on Music Information Retrieval (ISMIR), Late-Breaking/Demo session*, 2022.

[12] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[13] S. Doh, M. Won, K. Choi, and J. Nam, "Toward universal text-to-music retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[14] T. Chen, Y. Xie, S. Zhang, S. Huang, H. Zhou, and J. Li, "Learning music sequence representation from text supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[15] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, jan 2020.

[17] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800gb dataset of diverse text for language modeling," 2020.

[18] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[19] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank *et al.*, "Noise2music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[20] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd-workers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.

[21] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[22] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[23] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *Proceedings of*

the Advances in neural information processing systems (NeurIPS), 2022.

[24] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proceedings of the Advances in neural information processing systems (NeurIPS)*, 2017.

[25] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, 2023.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 2002.

[27] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004.

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations (ICLR)*, 2020.

[30] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[31] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *International Conference on Music Information Retrieval (ISMIR)*, 2009.

[32] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *International Conference on Music Information Retrieval (ISMIR)*, 2011.

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[34] Y. Zhang, J. Jiang, G. Xia, and S. Dixon, "Interpreting song lyrics with an audio-informed pre-trained language model," in *International Conference on Music Information Retrieval (ISMIR)*, 2022.

[35] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[36] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[37] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: a survey on deep learning-based image captioning," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[38] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal metric learning for tag-based music retrieval," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.