LARGE-LANGUAGE-MODEL-BASED CAPTION AUGMENTATION FOR LANGUAGE-QUERIED AUDIO SOURCE SEPARATION

Yoonah Song[1*], *Do Hyun Lee*[1*], *and Hong Kook Kim*[1,2,3]

[1] AI Graduate School, [2] School of EECS GIST, Gwangju 61005, Republic of Korea
[3] Aunion AI, Co. Ltd., Gwangju 61005, Republic of Korea
{yyaass0531@gm., zerolee12@gm., hongkook@}gist.ac.kr

ABSTRACT

This paper proposes a prompt-engineering-based caption augmentation approach for enhancing the performance of language-queried audio source separation (LASS) models. In the context of LASS, when large language models (LLMs) are utilized to generate augmented captions for audio clip descriptions, the choice of LLM prompts significantly influences the performance of LASS models. Hence, this study compares the performance of a LASS model using a dataset-dependent prompt (DDP) and a dataset-independent prompt (DIP). Experimental results on a small-sized benchmarking dataset reveal that the DDP-based caption augmentation approach achieves better speech quality than the corresponding DIP approach. However, not all DDP-generated captions guarantee quality improvement of the LASS models. Thus, a criterion is proposed to exclusively select effective captions based on their Bidirectional Encoder Representations from Transformers (BERT) similarity scores relative to the original caption. Subsequently, augmented captions with BERT similarity scores exceeding a predefined threshold are adopted for model training. The effectiveness of the proposed prompt-engineering-based approach is then evaluated on the baseline LASS model of DCASE 2024 Challenge Task 9. Performance evaluation results show that the baseline LASS model using the proposed prompt-generated caption outperforms the model using the original caption. The proposed prompt-engineering approach is also applied to AudioSep, a state-of-the-art model, to verify its validity across diverse LASS models. Ablation studies reveal that selecting appropriate prompts for LLM-based caption augmentation significantly enhances LASS performance. Furthermore, selective augmentation based on BERT similarity scores can further boost audio separation quality.

Index Terms— Language-queried audio source separation (LASS), large language model (LLM), caption augmentation, BERT similarity score, DCASE 2024 Challenge Task 9

1. INTRODUCTION

Source separation refers to the technique of isolating specific sound sources from a mixture of audio signals. Traditionally, this domain has primarily focused on tasks with predefined target sources, including speech enhancement [1], speech separation [2], and music source separation [3]. Recently, significant research efforts have been devoted toward universal sound separation [4], which seeks to segregate diverse real-world sound classes. However, owing to the vast number of possible sound sources, predefining all potential classes is nearly impossible.

To address this, researchers have explored a query-based sound separation approach utilizing visual [5] or audio queries [6] to separate specific sound sources. One such approach is language-queried audio source separation (LASS) [7], [8], which leverages natural language queries to identify and separate target sound sources. However, significant challenges related to data availability and quality hinder the training of deep learning models for LASS. Furthermore, the effectiveness of the language-query approaches relies on the user of the LASS model. This implies that the manner in which an audio clip is queried can vary widely depending on the time, location, and occasions of using the LASS model, even when the same user is involved. Consequently, annotating individual audio clips with inputs from numerous people is essential [9]. However, this annotation process is expensive and time-consuming, resulting in a limited number of captions for each audio clip. To address this data scarcity, the most intuitive solution is to utilize text augmentation techniques.

Notably, text augmentation research in the natural language processing (NLP) domain aims to improve the robustness of NLP models by generating diverse yet meaningful variations of original sentences. Easy Data Augmentation [10], a representative example of text augmentation approaches, adopts four techniques: synonym replacement, random insertion, random swap, and random deletion. These techniques enhance the robustness of text classifiers through text augmentation. Beyond textual content, audio and video captioning tasks aim to convert non-textual media into descriptive language, thus enhancing the accessibility and comprehension of audiovisual content. While audio captioning tasks generate textual descriptions of sound content [11], video captioning tasks automatically generate textual descriptions of actions and events depicted in videos [12]. These tasks, similar to LASS, involve handling both textual information and audiovisual content. However, current captioning research, such as [11] and [12], predominantly focuses on augmenting audio and video features to address the challenges related to data scarcity. Consequently, attempts to augment a linguistic expressions remain scarce.

Unlike traditional multimodal data augmentation approaches, which focus on diversifying audio and video content, our research focuses on augmenting data to enrich textual diversity. Specifically, by augmenting captions, our approach enables the LASS models to identify and utilize diverse captions conveying the same meaning. For example, identifying "a sound of thin plastic rattling"

[*] Y. Song and D. H. Lee equally contributed.

[**] This work was supported in part by the GIST-MIT Research Collaboration grant and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2024-00411137). It was also supported by "Project for Science and Technology Opens the Future of the Region" program through the Innopolis Foundation funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312).

Table 1: Summary of the training datasets used in this paper.

Dataset	Data subset	# of Clips	# of Captions
FSD50K		40,966	40,966
Clotho v2		3,839	19,195
WavCaps	BBC	31,201	31,201
	SoundBible	1,232	1,232
	AudioSet	108,317	108,317

as “fire crackling.” Additionally, utilizing multiple captions for each audio clip enhances LASS performance. To further enhance the performance of LASS models, this study utilizes a large language model (LLM) to augment the caption of audio clips. Recently, numerous studies have developed approaches for text augmentation using LLMs, highlighting the significant impacts of LLM input prompts on the output quality [13]. Considering this, the current study investigates sophisticated prompt designs to enhance LLM-based caption augmentation. The key contributions of our research are summarized as follows:

- We first investigate how to effectively design an input prompt for augmenting captions using an LLM. Our results indicate that a dataset-dependent prompt (DDP), which is designed considering various sentence structures across different datasets, performs better than a dataset-independent prompt (DIP), which uses a single prompt regardless of the dataset.
- Given that all generated captions by LLM may not necessarily improve the training of LASS models, we establish a criterion for selecting captions. This criterion adopts the BERT similarity score to quantify the similarity between original and augmented captions. Subsequently, performance evaluations of the LASS model are conducted by selecting captions depending on their similarity scores. Our findings show that utilizing descriptive captions with a diverse range of similarity scores is more effective than focusing solely on those with high similarity to the original ones.
- We examine the effectiveness of the proposed prompts and selection criterion across different LASS models. The results reveal that the proposed approach demonstrates effective for the baseline model of DCASE 2024 Challenge Task 4 and the AudioSep model [8]. Consequently, the LASS model employing our caption augmentation approach is ranked first in the evaluation of DCASE 2024 Challenge Task 9.

The remainder of this paper is organized as follows. Section 2 describes the datasets used for the LASS model. Section 3 presents the LASS model and proposes our caption augmentation approach using an LLM-based prompt-engineering strategy. Section 4 presents a performance evaluation of the LASS model on the validation dataset of DCASE 2024 Challenge Task 9. Finally, Section 5 concludes this paper.

2. DATASET

The baseline system of DCASE 2024 Challenge Task 9 [7], [8] was developed using audio samples sourced from Clotho v2 [9] and Freesound Dataset 50K (FSD50K) [14]. Notably, all audio clips within these datasets were acquired from the Freesound platform. In FSD50K, captions for each audio clip were initially obtained by refining raw descriptions using ChatGPT. Meanwhile, captions in Clotho v2 were crowdsourced using annotators from English-speaking countries, resulting in five captions per audio clip. In addition to these datasets, WavCaps [15] dataset was also used as one of externally available datasets. The WavCaps [15] dataset included the Freesound subset, but the Freesound subset was excluded in this work. More detailed information regarding the selected datasets is outlined in Table 1.

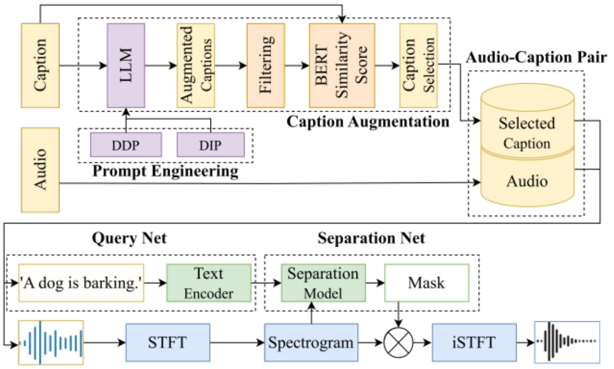

Figure 1: Procedure of training a LASS model using the proposed caption augmentation approach.

Table 2: Distribution of captions according to different datasets and prompts.

Dataset \ # of Captions		Original	DDP	DIP
FSD50K		162,485	40,966	30,991
WavCaps	BBC	31,201	131,969	27,225
	SoundBible	1,232	1,997	550
	AudioSet	108,319	406,139	75,300

In particular, FSD50K was an extensive dataset comprising 51,197 Freesound clips with human-labeled sound occurrences. Each audio clip was categorized based on 200 AudioSet ontology classes. Clotho v2 was an audio captioning dataset comprising 5,929 audio clips. Of these, 3,839 audio clips are allocated for development, 1,045 for validation, and 1,045 for evaluation. Each clip possessed five manually generated captions, varying in length from eight to twenty words. Finally, WavCaps in this study comprised 140,750 audio clips excluding the Freesound subset. Meanwhile, captions for these audio clips were generated by ChatGPT based on their raw audio descriptions. While consistent description-generation conditions were applied to SoundBible, and BBC, differing conditions were adopted for AudioSet.

3. PROPOSED LASS MODEL

We develop a LASS model based on the baseline model of DCASE 2024 Challenge Task 9 [7], [8]. This baseline LASS model comprises two key components: Query Net, which leverages Contrastive Language Audio Pretraining (CLAP) [16], and Separation Net which employs ResUNet [17]. Notably, the desired target source is conditioned by Query Net and separated by Separation Net. This paper proposes a prompt-engineering-based approach for LLM based caption augmentation, aiming to diversify the query representations of Query Net as if annotated by many humans.

Fig. 1 illustrates the training procedure of our LASS model using the proposed caption augmentation approach. Initially, a prompt corresponding to original caption of an audio clip is inputted into an LLM with prompt to generate multiple captions. The resulting captions are subsequently filtered to remove a prompt with the original sentences. The filtered captions are then incomplete or interrogative caption corresponding to an audio clip, compared to the original caption, in terms of their meaningfulness and

Table 3: Comparison of SDR, SDRi, and SI-SDR between DDP and DIP with randomly sampled 10,000 audio clips from different datasets.

Dataset	FSD50K				WavCaps							
					BBC+SoundBible				AudioSet			
Prompt Type	Original	DDP	DDP	DIP	Original	DDP	DDP	DIP	Original	DDP	DDP	DIP
# of Augmented Captions	-	39,586	7,498	7,498	-	41,359	8,578	8,578	-	37,634	6,879	6,879
SDR	3.960	4.432	4.237	4.113	4.577	4.968	4.807	4.764	4.465	5.038	4.808	4.764
SDRi	3.925	4.397	4.202	4.078	4.542	4.933	4.772	4.729	4.430	5.003	4.773	4.729
SI-SDR	0.293	1.470	0.852	0.618	1.083	2.431	2.090	1.961	1.092	2.311	2.215	1.848

Table 4: Comparison of SDR, SDRi, and SI-SDR in relation to BERT similarity scores for FSD50K dataset, using approximately 11,277 randomly selected augmented captions.

Threshold	0	0.700	0.850
# of Augmented Captions	11,277	11,277	11,277
SDR	4.236	4.369	4.167
SDRi	4.201	4.334	4.132
SI-SDR	1.195	1.372	1.080

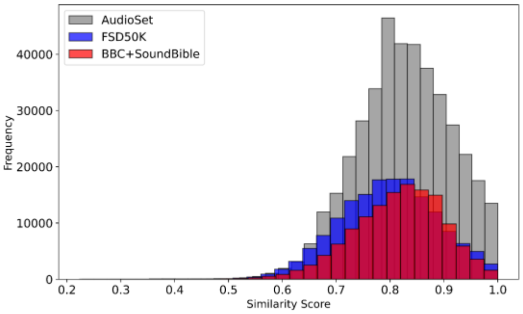

Figure 2: Distribution of BERT similarity scores for DDP-based augmented captions.

diversity, using BERT-based similarity scores [18]. Next, the selected captions are paired with their corresponding audio clips, forming multiple pairs of caption-audio clips by copying the original audio cli p to make the pairs. Finally, the LASS shown in the lower arm of the figure is trained using the augmented pairs of caption-audio clips. Note here that Microsoft's Phi-2.0 LLM [19] is used for caption generation, which is a 2.7 billion-parameter language model known for its superior comprehension and generation capabilities compared to the Llama-7B model.

3.1. Quality of augmented captions based on input prompts

We first investigated how to effectively design an input prompt for augmenting captions using the LLM. A review of relevant studies revealed that individual datasets require unique prompts customized to their specific attributes rather than general prompts [16]. Therefore, we hypothesized that crafting prompts tailored to the attributes of each dataset could enhance the quality of the generated captions. This approach led to the development of dataset-dependent prompts (DDPs), which generate sentences closely resembling original descriptions while meeting prompt requirements. On the other hand, the dataset-independent prompts (DIPs) were designed to be applicable even without prior information of individual dataset characteristics.

Second, we customized DDP based on the caption-generation conditions adopted in [9], [14], [15] curating a distinct prompt for each of the FSD50K, AudioSet subset, and BBC+SoundBible subset. Here, SoundBible+BBC means a subset datasets combining the BBC and SoundBible subsets because they share identical caption-generation conditions [15].

Next, to formulate a DIP, we initially referenced the caption-generation prompts utilized in WavCaps [15] and Clotho v2 [9]. However, they used a prompt to generate a sentence by only considering the event label. It was evident that these prompts might not be suitable for our study because we needed to augment sound description captions at the sentence level but not the event word level. To remedy this issue, we needed to redesign the DIP so that it could consider a sentence with a similar meaning to the original prompt. The detailed d esign process of the DIP is described in [20]. After generating the captions using the DDP or DIP, the captions were filtered and selected, as described above. In particular, the BERT similarity score between the original and each generated caption was computed. Then, captions whose scores were higher than a predefined threshold were selected, while captions with a similarity score of 1.0 was removed because it implied that the generated ca0ptions were identical to the original caption. Table 2 presents the number of captions augmented using each prompt for different datasets after applying the filtering process to the generated captions.

Table 3 compares the signal-to-distortion ratio (SDR) between the DDP and DIP. Due to the large data sizes, we randomly sampled 10, 000 audio clips from each dataset and used the corresponding caption-audio pairs for this experiment. While both types of prompts were designed to augment four captions per audio clip, the DIP produced approximately 1 to 1.5 captions on average, whereas the DDP consistently generated 4 captions on average. Specifically, the LASS model employing the DDP for FSD50K achieved an SDR of 4.432 dB, whereas that employing the DIP attained an SDR of 4.113 dB. Meanwhile, for the BBC+ SoundBible subset, the LASS models using the DDP and DIP achieved SDRs of 4.968 and 4.764 dB, respectively. Finally, the LASS models using the DDP and DIP achieved SDRs of 5.038 and 4.764 dB, respectively, for the AudioSet subset. As shown in the table, there was a similar tendency in other quality measures such as SDR improvement (SDRi) and scale-invariant (SI)-SDR, when comparing the DDP-based augmentation with DIP-based one. To ensure a fair quality comparison, we conducted additional experiments with equal numbers of augmented captions for both methods. Consequently, DDP consistently outperformed DIP in SDR and SI-SDR, indicating superior caption quality. It was also revealed that better performance was achieved with more captions for DDP.

Based on these results, it was proven that the DDP was more effective than the DIP, the LASS model employing DIP-based augmentation demonstrated performance improvement over that using original captions. Hence, in cases with limited knowledge regarding specific data characteristics, our DIP can be also useful as a viable alternative.

3.2. Quality of augmented captions based on the BERT similarity score

Since it is uncertain whether all captions automatically generated

Table 5: Performance comparison of different LASS models trained using various combinations of training datasets with/without caption augmentation on the validation dataset of DCASE 2024 Challenge Task 9.

Model	Training Dataset	Training Approach	Caption Augmentation	SDR	SDRi	SI-SDR
Baseline	Baseline Dev Set (FSD50K + Clotho v2)	Full	N/A	5.817	5.782	3.837
		Full	DIP	6.547	6.512	4.636
		Full	DDP	6.716	6.681	4.729
	Baseline Dev Set + WavCaps	Full	N/A	7.500	7.465	5.795
		Full	DIP	7.750	7.715	6.161
		Full	DDP	7.818	7.783	6.321
AudioSep	-	Pretrained	-	8.195	8.160	6.708
	Baseline Dev Set + WavCaps	Fine-tuning	N/A	8.370	8.335	7.109
		Fine-tuning	DIP	8.459	8.424	7.072
		Fine-tuning	DDP	8.489	8.454	7.198

by the LLM effectively contribute to the training of the LASS model, we establish a criterion for selecting captions. To this end, the BERT similarity score is used to measure the similarity between the original and augmented captions, because the BERT similarity score can assess the similarity of each token in the candidate sentence using contextual embeddings [18].

Fig. 2 depicts the distribution of BERT similarity scores for DDP-based augmented captions. Each distribution seems to be a Gaussian distribution with a mean of 0.85 and a little different variance. It was observed from the comparison between the original and generated captions that the generated captions with similarity scores below a certain threshold could be unsuitable as sound descriptions. For instance, the original caption “A musician plays a tune on a wind instrument” was augmented to “The sound of thunder fills the air, shaking the ground and captivating everyone's attention,” scoring 0.6, thus significantly differing in meaning.

Next, a performance evaluation of the LASS model on the FSD50K dataset was conducted by selecting captions depending on the BERT similarity score. Table 4 compares the objective performance of the LASS models trained by the selected captions according to different thresholds, where approximately 11,277 randomly selected captions were used for each threshold. As shown in the table, we set the threshold as 0.7 for caption selection, because using descriptive captions with a diverse range of similarity scores was more effective than using those with high similarity to the original ones.

4. PERFORMANCE EVALUATION

In this section, we evaluated the performance of the LASS models employing DDP and DIP. In addition to the baseline LASS model, the AudioSep model [8] was also trained to examine the effectiveness of the proposed prompts and selection criterion on different LASS models. Table 5 compares SDR, SDRi, and SI-SDR of different LASS models trained using various combinations of training datasets with/without caption augmentation on the validation dataset of DCASE 2024 Challenge Task 9. In this work, the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 64 was applied for 100 epochs to train the LASS models. Notice that the BERT similarity score threshold of selecting captions was all set to 0.7.

As shown in Table 5, the baseline LASS model trained on Baseline Dev Set achieved an SDR of 5.817 dB, which is consistent with the DCASE 2024 Challenge Task 9 baseline checkpoint [21], [22]. Augmenting the Baseline Dev Set with DIP-based generated captions increased the SDR to 6.547 dB, and DDP-based captions further improved it to 6.716 dB, demonstrating better performance compared to DIP-based ones. Training on the WavCaps dataset (excluding Freesound) resulted in an SDR of 7.500 dB, with DDP-based captions pushing the SDR to 7.818 dB.

Next, the pretrained AudioSep model, which was trained on over 2 million clips from weakly labeled datasets such as AudioSet, VGGSound, and AudioCaps, was utilized to validate the general applicability of the DDP-based caption generation approach. As shown in Table 5, the pretrained AudioSep model achieved an SDR of 8.195 dB, surpassing that of the baseline LASS model trained with DDP-based augmented captions. Fine-tuning this model using Baseline Dev Set and WavCaps dataset increased the SDR to 8.370 dB. On the other hand, the AudioSep model using DDP-based generated captions reached SDR of 8.489 dB, demonstrating performance compared to that using DIP-based captions. Thus, the AudioSep model fine- tuned by employing DDP-based caption augmentation demonstrated the best performance in terms of the SDR, SDRi, and SI-SDR.

5. CONCLUSION

To enhance the performance of LASS models, this paper proposed DDP-based caption augmentation as a means of prompt engineering. Specifically, two prompts were developed: a DDP, which is tailored to the characteristics of a specific dataset, and a DIP, which could be used without dataset information. Utilizing these prompts, five captions were generated for each audio clip using an LLM, following which selective learning of the augmented captions was performed based on BERT similarity scores. Subsequently, the SDR performance of the baseline and AudioSep models with DDP-based and DIP-based caption augmentation was assessed. Our findings demonstrated that the DDP, which dependently considered the unique characteristics of each dataset, yielded more suitable results. Furthermore, performance improvements were observed as the BERT similarity scores between the original and augmented captions reached values of 0.700 or higher. Collectively, these findings underscore the importance of customized prompt engineering in enhancing LASS performance through data augmentation.

In our study, we utilized LLM to augment captions to enhance the performance of the LASS model. While our approach showed improved results, several limitations should be noted. Primarily, the use of LLMs and the design of appropriate prompts for caption augmentation are still largely unexplored. Thus, our approach may not have fully leveraged the potential of the model. Additionally, although performance improvements were observed in LASS, it is uncertain if these enhancements can be generalized to other tasks using caption-audio paired data, such as audio captioning.

6. REFERENCES

[1] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” in *Proc. Interspeech*, 2017, pp. 3642–3646.

[2] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio Speech Lang.Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] S. Park, T. Kim, K. Lee, and N. Kwak, “Music source separation using stacked hourglass networks,” *arXiv preprint* arXiv:1805.08559, 2018.

[4] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, “Universal sound separation,” in *Proc. WASPAA*, 2019, pp. 175–179.

[5] R. Gao and K. Grauman, “VisualVoice: Audio-visual speech separation with cross-modal consistency,” in *Proc. CVPR*, 2021, pp. 15490–15500.

[6] B. Gfeller, D. Roblek, and M. Tagliasacchi, “One-shot conditional audio filtering of arbitrary sounds,” in *Proc. ICASSP*, 2021, pp. 501–505.

[7] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, “Separate what you describe: Language-queried audio source separation,” in *Proc. Interspeech*, 2022, pp.1801–1805.

[8] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, “Separate anything you describe,” *arXiv preprint,* arXiv:2308.05037, 2023.

[9] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” in *Proc. ICASSP*, 2020, pp. 736–740.

[10] J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” in *Proc. EMNLP-IJCNLP*, 2019, pp. 6382–6388.

[11] Z. Ye, Y. Wang, H. Wang, D. Yang and Y. Zou, “FeatureCut: An adaptive data augmentation for automated audio captioning,” in *Proc. APSIPA*, 2022, pp. 313–318

[12] C. Wang, H. Yang, and C. Meinel, “Image captioning with deep bidirectional LSTMs and multi-task learning,” *ACM TOMM*, vol. 14, issue 2s, pp. 1–20, 2018.

[13] Y. Zhou, A. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” in *Proc. ICLR*, 2023. Available: https://openreview.net/forum?id=92gvk82DE-.

[14] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “FSD50K: An open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio Speech Lang. Process,* vol. 30, pp. 829–852, 2022.

[15] 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.

[16] C. Li, M. Zhang, Q. Mei, W. Kong, and M. Bendersky, “Learning to rewrite prompts for personalized text generation,” in *Proc. ACM Web Conference*, 2024, pp. 3367–3378.

[17] Q. Kong, Y. Cao, H. Liu, K.Choi, and Y. Wang, “Decoupling magnitude and phase estimation with deep ResUNet for music source separation,” in *Proc. ISMIR*, 2021. Available: https://doi.org/10.48550/arXiv.2109.05418.

[18] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *Proc. ICLR*, 2020. Available: https://openreview.net/forum?id=SkeHuCVFDr.

[19] M. Javaheripi and S. Bubeck, “Phi-2: The surprising power of small language models.” Available at https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, and accessed on May 01, 2024.

[20] D. H. Lee, Y. Song, and H. K. Kim, “performance improvement of language-queried audio source separation based on caption augmentation from large language models for DCASE Challenge 2024 Task 9,” *arXiv preprint*, arXiv:2406.11248, 2024.

[21] L. Xubo and Z. Yan, “DCASE 2024 Task 9: Language-queried audio source separation | pre-trained weights for the baseline system.” Available at https://zenodo.org/records/10887460, and accessed on May 01, 2024.

[22] https://dcase.community/challenge2024/task-language-queried-audio-source-separation. Accessed on May 01, 2024.