A SOUND DESCRIPTION: EXPLORING PROMPT TEMPLATES AND CLASS DESCRIPTIONS TO ENHANCE ZERO-SHOT AUDIO CLASSIFICATION

Michel Olvera, Paraskevas Stamatiadis, Slim Essid

A SOUND DESCRIPTION: EXPLORING PROMPT TEMPLATES AND CLASS DESCRIPTIONS TO ENHANCE ZERO-SHOT AUDIO CLASSIFICATION

Michel Olvera, Paraskevas Stamatiadis, Slim Essid

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
{olvera, paraskevas.stamatiadis, slim.essid}@telecom-paris.fr

ABSTRACT

Audio-text models trained via contrastive learning offer a practical approach to perform audio classification through natural language prompts, such as "this is a sound of" followed by category names. In this work, we explore alternative prompt templates for zero-shot audio classification, demonstrating the existence of higher-performing options. First, we find that the formatting of the prompts significantly affects performance so that simply prompting the models with properly formatted class labels performs competitively with optimized prompt templates and even prompt ensembling. Moreover, we look into complementing class labels by audio-centric descriptions. By leveraging large language models, we generate textual descriptions that prioritize acoustic features of sound events to disambiguate between classes, without extensive prompt engineering. We show that prompting with class descriptions leads to state-of-the-art results in zero-shot audio classification across major ambient sound datasets. Remarkably, this method requires no additional training and remains fully zero-shot.

Index Terms— Zero-shot audio classification, audio-text models, contrastive language-audio pretraining, in-context learning

1. INTRODUCTION

Multimodal contrastive pretraining has been used to train multimodal representation models on large amounts of paired data. This approach leverages contrastive learning to align representations across different modalities, promoting a shared embedding space that improves semantic understanding across modalities. Examples include Contrastive Language-Image Pretraining (CLIP) [1], which aligns visual and textual representations, and the more recent Contrastive Language-Audio Pretraining (CLAP), which extends these principles to align audio and textual representations [2, 3, 4, 5].

Following pretraining, CLAP exhibits a well-structured feature space, yielding robust, general-purpose representations well-suited for downstream training. Moreover, it also demonstrates exceptional transferability as evidenced by its impressive zero-shot performance across classification, captioning, retrieval, and generation tasks [3, 6, 7].

Extensive research on CLIP has revealed that classification scores are significantly influenced by alterations in prompt formulation and language nuances. For instance, varying the description of a concept, using synonyms, or modifying the grammatical structure or wording, substantially affects performance outcomes [8, 9, 10]. Besides, prompts offering more context or specificity tend to yield more accurate results [11, 12, 13].

Similarly, CLAP inherits sensitivity to prompting from its contrastive pretraining approach. Yet, the systematic exploration of prompt robustness in CLAP remains limited, despite few works highlighting the sensitivity of classification to prompt variations [14, 15]. These works, primarily conducted on the ESC50 dataset and limited to up to five prompt templates, shed initial light on these variations. However, robustness to prompt changes is likely to vary across different datasets. Addressing this gap, recent efforts have explored alternative approaches, such as prompt tuning strategies and lightweight adapters, to mitigate the reliance on manually engineered prompts [16, 17] with an explicit focus on adapting CLAP to downstream tasks or new domains.

In this work, we propose a tuning-free approach that prompts CLAP models with descriptions of class labels to enhance zero-shot audio classification. While using keywords such as "audio," "hear," and "sound" in prompt templates primes the text encoder to focus on audio-related concepts, we hypothesize that enriching prompts with explicit class descriptions can further enhance the model's ability to clarify the meaning of class labels, particularly in scenarios where labels are ambiguous. Ambiguity stems from both the textual and audio aspects of the data. Textual ambiguity arises from homonyms, where words possess multiple meanings, and from the lack of contextual clues (*e.g.*, "bat" as both an animal and sports equipment). On the audio side, ambiguity arises from acoustically similar sound categories, such as distinguishing between bird vocalizations (*e.g.*, raven vs. crow calls) and musical instruments (*e.g.*, violin vs. viola). Thus, detailed prompts may clarify sounds heavily reliant on context, and help disambiguate acoustically similar sounds. Such descriptions can also disambiguate abstract sounds such as "white noise" and compensate for knowledge gaps or limited exposure to certain terms. For instance, clarifying "Geiger counter", as "a detection device that clicks or beeps when detecting radiation" could improve correlations of audio and text features.

To validate our hypothesis, we leverage Large Language Models (LLMs) for their knowledge of sound semantics. Specifically, we used Mistral[1] to describe the acoustic properties of class labels. Our study demonstrates that using audio-centric descriptions of class labels as prompts helps CLAP better ground acoustic features with semantic descriptions, significantly boosting zero-shot classification scores across major environmental sound datasets. Remarkably, our method even outperforms learnable prompt strategies, all without the need for additional training, while remaining entirely zero-shot.

2. METHODOLOGY

We first describe the zero-shot audio classification task, then our adaptive class selection strategy and finally we motivate our LLM-generated class descriptions.

This work was supported by the Audible project, funded by French BPI.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

Class	Base	Context	Ontology
Mandolin	A stringed musical instrument, played with a plectrum, characterized by its small size, high-pitched sound, and distinctive twang.	A stringed musical instrument with a distinctive, twangy sound, often associated with folk or bluegrass music. Typically played by plucking or strumming the strings, producing a bright, melodic tone.	A stringed musical instrument with a distinctive, twangy sound, often used in folk and pop music.
Rail transport	The sound of trains moving on rails, characterized by the clacking of wheels and the rumbling of engines.	The sound of trains moving along rails, characterized by a steady, rhythmic clacking or clicking noise. Often heard in urban or rural areas with rail infrastructure.	The rumbling and clanking sounds produced by trains moving on rails, characterized by their speed and intensity, classified under transportation-related sounds.
Toot	A short, high-pitched sound produced by blowing air through a small opening, often used as a signal or warning.	A short, sharp sound, typically produced by blowing air through a small opening, such as a whistle or a musical instrument.	A short, high-pitched sound produced by a whistle or other musical instrument, often used as a signal or warning.
Stream	A continuous flow of water or other liquid, often characterized by its sound as it flows over rocks or other obstacles.	A continuous flow of water, often heard in natural environments like rivers, lakes, or waterfalls, characterized by the sound of water flowing over rocks or other surfaces.	A continuous flow of sound, often characterized by its rhythmic patterns and timbre, belonging to the category of natural environmental sounds.

Table 1: Example descriptions of randomly sampled class labels from the datasets considered in this work, generated with Mistral-7B [18].

2.1. Zero-shot audio classification

Given a set of target categories C and a query audio sample a, the zero-shot audio classification protocol in CLAP defines the classification problem as a nearest neighbor retrieval task. The predicted category $\hat{c}$ is determined as follows:

$$\hat{c} = \arg\max_{c \in C} \text{sim}(\phi_{\text{A}}(a), \phi_{\text{T}}(c)), \tag{1}$$

where C represents the set of class labels, a denotes the input audio, and ϕ_{A} and ϕ_{T} are the audio and text encoders, respectively. The function $\text{sim}(\cdot, \cdot)$ corresponds to the similarity metric, typically the cosine similarity.

To enhance zero-shot audio classification, we propose using both class labels and their descriptions to resolve ambiguities. Given a set of target categories C, definitions D, the predicted category $\tilde{c}$ is determined by:

$$\tilde{c} = \arg\max_{c \in C} \text{sim}(\phi_{\text{A}}(a), \phi_{\text{T}}(c + d_c))), \tag{2}$$

where $d_c \in D$ is the description corresponding to class c, and the $+$ operator denotes the textual combination of the class label c and its description d_c.

2.2. Adaptive class description selection

We devise an adaptive strategy that incorporates descriptions selectively for classes potentially ambiguous to the text encoder. Let $\text{P}_{\text{class-only}}$ and $\text{P}_{\text{class-description}}$ represent the classification performance for class c using setups involving classes only or classes with descriptions as in Equations (1) and (2), respectively. We decide for class c which setup to apply through the decision function $\text{M}(c)$:

$$\text{M}(c) = \begin{cases} \hat{c} & \text{if } \text{P}_{\text{class-only}} \geq \text{P}_{\text{class-description}} \\ \tilde{c} & \text{if } \text{P}_{\text{class-description}} > \text{P}_{\text{class-only}}. \end{cases} \tag{3}$$

The function $\text{M}(c)$ decides whether a class should include a description based on cross-validation of results.

2.3. Generation of audio-centric descriptions with LLMs

Given audio event class labels, we propose to use Large Language Models (LLMs) to generate audio-centric descriptions for them automatically, as manual collection of descriptions entails a labor-intensive endeavor. LLMs, trained on vast text data, have a deep understanding of language, which we exploit for their knowledge of sound semantics. Our method, adapted from [19], involves three steps. First, we provide a general description of the task. Second, we combine these instructions with in-context demonstrations, including a few paired label-description examples. Finally, we provide the LLM with the class labels, heuristic constraints, and specific output format details to generate audio-centric descriptions.

Using this method, we generated three types of descriptions: *base descriptions*, *context-aware descriptions*, and *ontology-aware descriptions*. All are audio-centric. *Base descriptions* reflect the acoustic properties and characteristic sounds of the class labels. *Context-aware* descriptions add details about the typical locations and circumstances of encountering the sounds, including the physical environment, associated objects, and the function of the sound within its context. *Ontology-aware* descriptions capture the acoustic properties and characteristic sounds of each class label while also considering their relationships with coarse high-level concepts. Table 1 provides a few examples of the generated descriptions. The complete list of class descriptions and the prompts used to generate them are available on our companion website.[2]

3. EXPERIMENTAL SETUP

We detail our experimental approach, including model and dataset selection, evaluation metrics, and experiments to explore different prompt strategies and their impact on classification.

3.1. Models

We adopt two state-of-the-art audio-text models pre-trained via contrastive learning, namely LAION-CLAP (LA) and Microsoft CLAP 2023 (MS). The former utilizes RoBERTa [20] as its text encoder, while the latter leverages GPT-2 [21]. Both models rely on HTS-AT [22] as their audio encoder.

3.2. Datasets and evaluation metrics

Downstream datasets. We select six major environmental sound datasets tailored for either single-class or multi-label classification. These include: **ESC50** [23], which contains 50 environmental sound classes with 2k labeled samples of 5 seconds each; **US8K** [24], comprising 10 urban sound classes and 8k labeled sound excerpts of 4 seconds each; **TUT2017** [25], consisting of 15 acoustic scenes classes and 52k files of 10 seconds each; **FSD50K** [26], featuring 51K audio clips of variable length (from 0.3 to 30 seconds each) curated from Freesound and comprising 200 classes;

[2] https://github.com/tpt-adasp/a-sound-description

AudioSet [27], a large-scale dataset encompassing 527 classes, with over 2 million human-labeled sound clips of 10 seconds from YouTube videos; and **DCASE17-T4** [25], a subset of AudioSet focused on 17 classes related to warning and vehicle sounds, containing 30k audio clips of 10 seconds each.

Evaluation setup and metrics. In our evaluation we consider all available splits (train/val/test) or folds, except for AudioSet, where only the test set was used. Note that some datasets do not allow for a fully zero-shot approach, as some audio files used in the evaluation were part of the pretraining data of the considered frozen CLAP models (*e.g.*, AudioSet and FSD50K). We believe that it is still interesting to analyse the corresponding results, bearing this fact in mind during the discussion. We use accuracy as the metric for single-class classification datasets (ESC50, US8K and TUT2017) and mean Average Precision (mAP) for multi-label classification datasets (FSD50K, AudioSet and DCASE17-T4). For experiments involving class-specific descriptions, a 5-fold cross-validation setting is employed. These folds were constructed on the data considered for evaluation *i.e.,* all splits/folds for all datasets, except for AudioSet where the test set is used. In this approach, training folds are used to derive the mapping M from Equation (3), while test folds are used to assess its generalization. Directly evaluating the mapping without cross-validation would yield overly optimistic results due to overfitting.

3.3. Zero-shot audio classification experiments

Prompting with class labels only We explore zero-shot audio classification using prompts with sanitized class labels (*i.e.*, replacing underscores in original labels with spaces, *e.g.*, *dog_barking* becomes *dog barking*). This is motivated by the fact that in our early experiments we observed that this strategy performs competitively compared to prompting with "This is a sound of", which has been preferred in the literature [14, 4]. Here, we systematically study the impact of using only class labels as prompts on classification performance. We examine four different formats to construct the start and end of a prompt: uppercase with a period (*e.g., Dog barking.*), uppercase without a period (*e.g., Dog barking*), lowercase with a period (*e.g., dog barking.*), and lowercase without a period (*e.g., dog barking*). The format yielding the highest performance for each model, termed as CLS, was selected as a reference for subsequent experiments involving class descriptions.

Prompting with templates. Inspired from CLIP [1], we explore a set of prompt templates as plausible alternatives to "This is a sound of", all tailored for the zero-shot audio classification task. We curated a set of 33 distinct prompts, drawing some from prior studies [14, 4, 15]. Our objective is to systematically evaluate the performance of these alternative prompts and their ensemble across multiple datasets. Each prompt follows the format *Template + class label*, *e.g.*, "A sound clip of dog barking.". We thus analyse the performance of three prompt configurations: $\text{PT}_{\text{Baseline}}$: The baseline prompt template "This is a sound of". PT_{Best}: The most effective prompt template identified among the 33 manually crafted alternatives. $\text{PT}_{\text{Ensemble}}$: Ensembling text embeddings from all considered prompt templates. Each prompt template begins with an uppercase letter and concludes with a period.

Prompting with class-specific descriptions. We investigate the impact of combining class labels and their descriptions generated by LLMs. The experimental setups include: CLS: Class label only. CD_{Base}: Audio-centric definitions generated by Mistral. $\text{CD}_{\text{Context}}$[3]: Context-aware descriptions. $\text{CD}_{\text{Ontology}}$: Ontological information related to the class label. $\text{CD}_{\text{Dictionary}}$: Definitions (non audio-centric) sourced from the Cambridge Dictionary of English.[4]

4. RESULTS AND DISCUSSION

In this section, we present and discuss the outcomes of our experiments, shedding light on the impact of various prompting strategies and the role of class descriptions in classification performance.

4.1. Sensitivity to prompt format

In Table 2, we report the average classification results across all evaluation datasets to examine the sensitivity of zero-shot classification performance to subtle variations in the input prompt format. We see surprising differences in performance due to minor alterations such as capitalization and punctuation, consistent with findings in [15]. A recent work on LLM behavior confirm that these seemingly minor changes in prompt format influence the model's internal representations, leading to distinct transformations within the embedding space that alter the output probability distribution in ways that affect classification performance [28]. We observe that, for both models, prompt variations in punctuation, irrespective of capitalization, significantly affect performance more than variations in capitalization without punctuation. Notably, the performance gap between the most and least effective formats was 5.46% for model LA and 8% for model MS, pointing out how critical it is to select an optimal format to maximize classification scores. Consequently, subsequent experiments adopted the best-performing format for each model.

Prompt format	Model	
	LA	MS
class label (*e.g., dog barking*)	0.5059	0.5256
class label. (*e.g., dog barking.*)	0.5524	**0.5735**
Class label (*e.g., Dog barking*)	0.5110	0.49344
Class label. (*e.g., Dog barking.*)	**0.5605**	0.5395

Table 2: Average model performance scores across all datasets for different input prompt formats.

4.2. Comparison of prompting strategies

In Table 3, top-panel, we show results that assess the impact on classification performance when prompting CLAP models using only the class label and various prompt templates and an ensemble of these prompts. Our findings reveal that using the class label alone (CLS) often yields superior performance compared to the prompt template "This is a sound of" ($\text{PT}_{\text{Baseline}}$). Specifically, CLS demonstrates better results than $\text{PT}_{\text{Baseline}}$ on the majority of datasets, with model MS showing an absolute improvement of 1.07%. However, for model LA, CLS showed a slight underperformance of 0.67%, largely due to lower scores on the TUT2017 and DCASE17-T4 datasets.

[3]We did not consider context-aware descriptions for TUT2017 because these were very similar to base descriptions. Unlike other datasets, TUT2017 comprises labels that refer to acoustic scenes. This explains the similarity, as both type of descriptions indicate context.

[4]When definitions where not available in the Cambridge Dictionary, definitions were sourced from WordNet, Wikipedia, and FreeBase.

Method	ESC50		US8K		TUT2017		DCASE17-T4		FSD50K		AudioSet		Average	
	LA	MS	LA	MS	LA	MS	LA	MS	LA	MS	LA	MS	LA	MS
CLS	0.9280	0.9280	0.7980	0.8737	0.4242	0.5717	0.4443	0.3772	0.5409	0.5137	0.2277	0.1764	0.5605	0.5735
$PT_{Baseline}$	0.915	0.893	0.7747	0.7855	0.4890	0.4547	0.4670	0.4674	0.5308	0.5052	0.2269	0.2708	0.5672	0.5628
PT_{Best}	0.9415	0.9585	0.8133	0.8624	0.5041	0.6192	0.5220	0.4583	0.5765	0.5372	0.2855	0.2708	0.6071	0.6176
$PT_{Ensemble}$	0.9295	0.95	0.7893	0.8506	0.4944	0.6111	0.4851	0.4075	0.5744	0.5424	0.2560	0.2063	0.5881	0.5946
Adaptive class description selection (mean scores across five folds)														
$CD_{Dictionary}$	0.9535	0.9205	0.8632	0.8891	0.5770	0.5630	0.4704	0.3776	0.5623	0.4972	0.2727	0.1924	0.6165	0.5733
CD_{Base}	0.9480	0.9505	0.8336	0.8926	0.5790	**0.6219**	0.4705	0.3911	0.5654	0.5039	0.2803	0.1963	0.6128	0.5927
$CD_{Context}$	0.9455	0.9595	0.8597	0.8782	-	-	**0.4742**	0.3801	**0.5720**	0.5128	**0.2891**	0.2022	**0.6281**	0.5865
$CD_{Ontology}$	0.9495	**0.9635**	0.8480	**0.9017**	0.5030	0.5670	0.4589	0.3748	0.5676	0.5074	0.2830	0.1998	0.6017	0.5857
CD_{All}	0.9491	0.9485	0.8511	0.8904	0.5530	0.5840	0.4685	0.3809	0.5668	0.5053	0.2813	0.1976	0.6142	0.5845
SOTA	0.96 [15]		0.8526 [17]		0.5438 [17]		-		0.52 [15]		0.102 [17]		-	

Table 3: Zero-shot classification scores across 6 downstream tasks. Evaluation metrics: Accuracy for ESC50, US8K and TUT2017; mean Average Precision (mAP) for DCASE17-T4, FSD50K and AudioSet.

We report the best-performing prompt template, PT_{Best}, among those considered as plausible alternatives to $PT_{Baseline}$ for each dataset. On average, PT_{Best} outperformed $PT_{Baseline}$, with an absolute improvement of 3.99% and 5.48% for LA and MS, respectively. The relevance of this result brings to light the existence of better manually crafted prompt templates than *This is a sound of*. Table 4 lists the best-performing prompt template for each evaluation dataset. Interestingly, the absence of a "universal" template calls for customization to specific datasets and models to optimize performance, given that certain templates may align better with particular dataset labels. Additionally, prompt ensembling ($PT_{Ensemble}$) outperformed individual prompts like CLS and $PT_{Baseline}$, but did not exceed PT_{Best}, which can be attributed to less effective prompts in the ensemble, potentially diminishing its overall efficacy.

Dataset	Models	
	LA	MS
ESC50	*Listen to*	*A recording of*
US8K	*I can hear*	*Listen to an audio of*
TUT2017	*This is a sound track of*	*Listen to an audio recording of*
DCASE17-T4	*A sound clip of*	*This is a sound of*
FSD50K	*A sound recording of*	*This is*
AudioSet	*This is an audio clip of*	*This is a sound of*

Table 4: Best-performing prompt templates per dataset.

4.3. Impact of class-specific descriptions

In Table 3, middle-panel, we assess the impact of class-specific descriptions on classification performance through our adaptive selection strategy, which determines which classes benefit from explicit descriptions. Our findings indicate that introducing class descriptions is indeed beneficial for disambiguating difficult classes, with audio-centric descriptions generally outperforming dictionary definitions. Focusing on model LA, class descriptions with contextual information ($CD_{Context}$) yielded the best results on average. While model MS also benefited from class-specific descriptions, it showed modest gains across datasets, likely due to its pretraining on a larger volume of data, including more audio-caption pairs. For model MS, base audio-centric descriptions of class labels CD_{Base} were the most effective, but still could not outperform prompt template-based methods in the top-panel for datasets such as DCASE17-T4, FSD50k and AudioSet. However, our adaptive strategy incorporating all types of descriptions (CD_{All}) did not generalize as effectively compared to individual setups, which was somewhat disappointing.

A comparison with state-of-the-art zero-shot audio classification scores reported in the literature, as shown in bottom line of Table 3, reveals that our approach outperforms these benchmarks, including those utilizing prompt tuning strategies such as [17], across all evaluated datasets. The improvements are particularly notable for the US8K, TUT2017, FSD50K, and AudioSet datasets.

4.4. Disambiguation of classes through descriptions

In Table 5, we show the top-3 classes with the greatest absolute improvement in classification using base descriptions compared to the simple use of class labels for the AudioSet and FSD50K datasets. We observe some words are ambiguous in meaning, for which an explicit description is beneficial as indicated by the large absolute improvements. A full list of relative improvements for all datasets is available on our companion website.

Dataset	Class label	Δ Improvement [%]
AudioSet	*Bagpipes*	+40.12
	Fire engine, fire truck (siren)	+39.79
	Gargling	+36.34
FSD50K	*Fowl*	+67.75
	Scratching (performance technique)	+67.21
	Purr	+60.49

Table 5: Top-3 classes with highest absolute improved classification for model MS on AudioSet and FSD50K datasets using base audio-centric descriptions.

5. CONCLUSION

We demonstrated that prompt templates and class-specific descriptions can significantly impact the performance of zero-shot audio classification. While simple class labels can be highly effective, carefully crafted prompt templates and context-aware descriptions offer substantial improvements. Our findings advocate for a nuanced approach to prompt engineering, where the choice of format, content, and contextual information are tailored to the specific requirements of the model and dataset. Future work could explore automated methods for generating optimal prompts and descriptions, to further boost zero-shot audio classification scores.

6. REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*. PMLR, 2021, pp. 8748–8763.

[2] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. ICASSP*. IEEE, 2022, pp. 976–980.

[3] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. ICASSP*. IEEE, 2024, pp. 336–340.

[4] 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 *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[5] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," *arXiv preprint arXiv:2208.12208*, 2022.

[6] 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.

[7] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[8] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *Proc. ICLR*, 2023.

[9] B. An, S. Zhu, M.-A. Panaitescu-Liess, C. K. Mummadi, and F. Huang, "Perceptionclip: Visual classification by inferring and conditioning on contexts," in *Proc. ICLR*, 2024.

[10] A. Salinas and F. Morstatter, "The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance," *arXiv preprint arXiv:2401.03729*, 2024.

[11] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proc. ICCV*, 2023, pp. 15 691–15 701.

[12] K. Roth, J. M. Kim, A. Koepke, O. Vinyals, C. Schmid, and Z. Akata, "Waffling around for performance: Visual classification with random words and broad concepts," in *In Proc. of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 746–15 757.

[13] M. J. Mirza, L. Karlinsky, W. Lin, H. Possegger, M. Kozinski, R. Feris, and H. Bischof, "Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[15] S. S. Kushwaha and M. Fuentes, "A multimodal prototypical approach for unsupervised sound classification."

[16] Y. Li, X. Wang, and H. Liu, "Audio-free prompt tuning for language-audio models," in *Proc. ICASSP*. IEEE, 2024, pp. 491–495.

[17] S. Deshmukh, R. Singh, and B. Raj, "Domain adaptation for contrastive audio-language models," *arXiv e-prints*, pp. arXiv–2402, 2024.

[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[19] A.-M. Oncescu, J. F. Henriques, A. Zisserman, S. Albanie, and A. S. Koepke, "A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval," in *Proc. ICASSP*. IEEE, 2024, pp. 7300–7304.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[22] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. ICASSP*. IEEE, 2022, pp. 646–650.

[23] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM-MM*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[24] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM-MM*, 2014, pp. 1041–1044.

[25] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *Proc. IEEE/ACM Trans. Audio Speech Lang.*, vol. 27, no. 6, pp. 992–1006, 2019.

[26] 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.*, vol. 30, pp. 829–852, 2021.

[27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[28] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, "Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting," in *Proc. ICLR*, 2024.