What Do Language Models Hear?
Probing for Auditory Representations in Language Models

Jerry Ngo **Yoon Kim**
Massachusetts Institute of Technology
ngop@mit.edu yoonkim@mit.edu

Abstract

This work explores whether language models encode meaningfully grounded representations of sounds of objects. We learn a linear probe that retrieves the correct text representation of an object given a snippet of audio related to that object, where the sound representation is given by a pretrained audio model. This probe is trained via a contrastive loss that pushes the language representations and sound representations of an object to be close to one another. After training, the probe is tested on its ability to generalize to objects that were not seen during training. Across different language models and audio models, we find that the probe generalization is above chance in many cases, indicating that despite being trained only on raw text, language models encode grounded knowledge of sounds for some objects.

1 Introduction

Despite being trained only on surface-form strings (i.e., without explicit grounding), language models (LMs) have been shown to learn representations of perceptual concepts that plausibly mirror the grounded, physical representations of those same concepts. Examples of such concepts that have been investigated so far in the literature include color (Abdou et al., 2021), direction (Patel and Pavlick, 2021), size (Zhang et al., 2020; Grand et al., 2022), geography (Konkol et al., 2017; Liétard et al., 2021; Faisal and Anastasopoulos, 2023; Chen et al., 2023), time (Gurnee and Tegmark, 2023), and even visual representations (Ilharco et al., 2021; Merullo et al., 2022; Li et al., 2023). The alignment between an LM's induced representation of a concept (e.g., the space of word embeddings for colors) and its physical (or human perception-like) representation (e.g., RGB space) has direct implications for how much explicit grounding is necessary for an LM to learn about the "real world" referred to by the textual data on which it was trained. And inasmuch as grounding may be relevant for meaning and understanding,[1] these findings also have indirect implications for whether LMs can acquire (some operationalization of) meaning through text-only training (Bender and Koller, 2020).

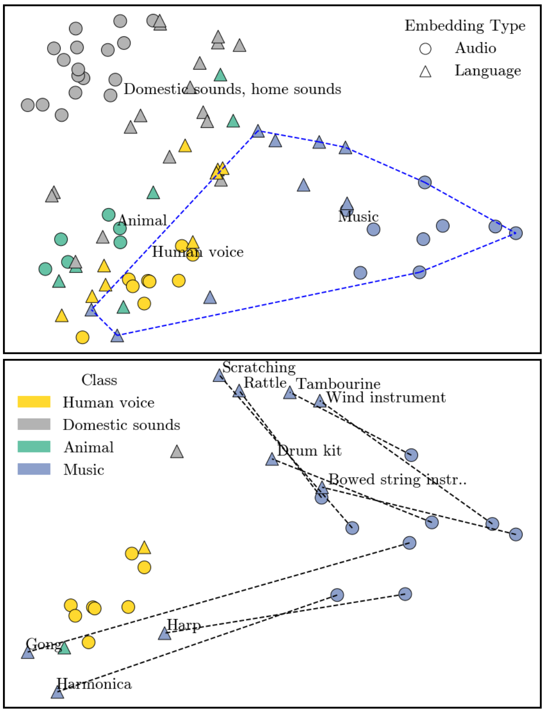


Figure 1: (Top) Language (triangle) and sound (circle) representations aligned via Procrustes analysis (Schönemann, 1966), visualized via PCA. The language representation is from BERT (Devlin et al., 2019) and the audio representation is from PaSST (Koutini et al., 2021). The classes are color-coded based on their parent nodes (i.e., `human voice`, `domestic sounds`, `animal`, `music`) according to the ontology from the FSD50K (Fonseca et al., 2021). (Bottom) A zoomed-in portion of the blue region of the top figure, which shows the structural similarities between the language and sound representations for the `music` category.

[1] See Pavlick (2023) and Søgaard (2023) for further discussion on the relationship between grounding and meaning.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5435–5448
August 11-16, 2024 ©2024 Association for Computational Linguistics

This work investigates the extent to which LMs encode perceptual representations of *sounds*. Past works have found that LM representations of some objects are partially isomorphic to representations of those same objects from vision models (Ilharco et al., 2021; Li et al., 2023), suggesting that LMs are able to learn nontrivial structures about the visual world through just text-only training. We extend this setup to sounds through the lens of probing (Belinkov, 2022), where we learn simple linear transformations that align the language representation for an object c to its sound representation (from a pretrained audio model). If this retrieval-based probe generalizes to objects that were not seen during training, this suggests that there are structural similarities between the language and sound representations, i.e., LMs have learned meaningfully grounded representations of c despite being just trained on raw text.

We conduct the sound probing study across 6 language models and 3 audio models. The language representations include those from word vector-only models (GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013)), encoders (BERT (Devlin et al., 2019), T5 (Raffel et al., 2020)), and decoders (GPT-2 (Radford et al., 2019), LLaMA (Touvron et al., 2023)). On the audio side, we experiment with two types of models: self-supervised models that have been pretrained without access to any external labels (AudioMAE; Huang et al., 2022), and supervised models that have been pretrained on sound event classification (finetuned AudoMAE, PANN (Kong et al., 2020), PaSST (Koutini et al., 2021)). While all audio models are trained without explicit access to symbolic language data, the representations from supervised models implicitly encode more human perception-like priors given that the classification task itself incorporates information about what snippet of sound constitutes a salient-enough signal to humans to warrant its being classified as a distinct event. That is, purely self-supervised models are more likely to encode more physical (i.e., acoustic) representations whereas supervised models are more likely to encode more human perception-like (i.e., auditory) representations.

On both acoustic- and auditory-like sound representations, we find that all language models generalize to unseen classes at an above-chance level. We also find that the generalization performance is typically better for sound representations that have been supervised on sound event classification.

2 Probing for Auditory Knowledge

2.1 Preliminary Study: Procrustes Analysis

We perform a preliminary qualitative study to test the feasibility of aligning language and sound representations. Let $\mathcal{C} = \{\texttt{car}, \texttt{bus}, \texttt{harmonica}, \texttt{harp} \ldots\}$ be a set of objects. Further let $f_{\text{LM}} : \Sigma^* \rightarrow \mathbb{R}^{d_1}$ be a text encoding function from a pretrained LM that produces a d_1-dimensional vector representation of a sentence (e.g., via averaging the contexualized word embeddings of c occurring in some sentence $x \in \Sigma^*$), and similarly let $f_{\text{AM}} : \mathbb{R}^{T \times m} \rightarrow \mathbb{R}^{d_2}$ be an audio encoding function from a pretrained audio model that takes in an m-dimensional audio signal[2] of (up to) length T and produces a d_2-dimensional vector representation of that audio signal. For each $c \in \mathcal{C}$, let $\text{text}(c) \in \Sigma^*$ be the text template for c that describes c's sound,[3] and further let $\text{sound}(c)$ be a sound associated with c (e.g., the sound of a harmonica if $c = \texttt{harmonica}$). We are interested in comparing the space of induced text representations $\mathcal{C}_{\text{language}} = \{f_{\text{LM}}(\text{text}(c)) : c \in \mathcal{C}\}$ with the sound representations $\mathcal{C}_{\text{sound}} = \{f_{\text{AM}}(\text{sound}(c)) : c \in \mathcal{C}\}$; if the geometry of these representations is "similar" in some way (e.g., they are isomorphic), then we can infer that the pretrained LM has nontrivial knowledge of sounds that are associated with objects in $\mathcal{C}$.[4]

As an initial study, we analyze these two spaces with Procrustes analysis. Let $\mathbf{C}_{\text{language}} \in \mathbb{R}^{|\mathcal{C}| \times d_1}$ be the matrix obtained by stacking the text representations for each $c \in \mathcal{C}$, and similarly for $\mathbf{C}_{\text{sound}} \in \mathbb{R}^{|\mathcal{C}| \times d_2}$.[5] Since we generally have $|\mathcal{C}| < d_2 < d_1$, we first perform PCA on both matrices to obtain $\mathbf{C}'_{\text{language}} \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ and $\mathbf{C}'_{\text{sound}} \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$. Procrustes analysis aligns these two matrices via minimizing

$$\min_{\mathbf{Q} \in \mathbb{O}^{|\mathcal{C}| \times |\mathcal{C}|}} \|\mathbf{C}'_{\text{language}}\mathbf{Q} - \mathbf{C}'_{\text{sound}}\|^2,$$

where $\mathbb{O}^{|\mathcal{C}| \times |\mathcal{C}|}$ is the set of orthogonal matrices, i.e., $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$. We perform PCA again to two dimensions and visualize the resulting representation space. Figure 1 shows this analysis for BERT

[2] Where each dimension corresponds to a particular frequency in the case of spectrograms.

[3] Our template is "`the sound of c`."

[4] This is the case only if the pretrained audio model is trained without access to any symbolic language. This will indeed be the case in our experiments.

[5] In our dataset, there are multiple audio snippets associated with a single c. We average all audio embeddings associated with c to obtain the object-level sound representation.

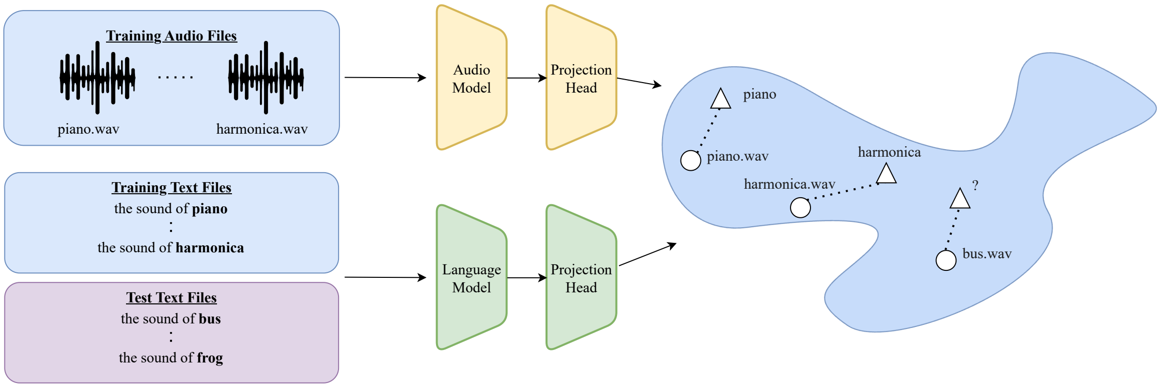


Figure 2: An overview of our experimental setup. We randomly split a set of classes into mutually exclusive train/test sets. On the training set (blue), we use a contrastive loss to learn linear transformations (i.e., projection heads) of the sound and language representations such that a language representation of a class is close in cosine distance to the sound representation of the same class. After training, we apply the learned probe on audio snippets of classes from the test set, and retrieve the most similar text representation (from classes in both the train and test sets). We then test whether the retrieved class corresponds to the actual class.

(Devlin et al., 2019) and PaSST (Koutini et al., 2021) on the FSD50K (Fonseca et al., 2021) dataset. While far from perfectly aligned, there are reasonable symmetries. This motivates a more controlled, quantitative study described below.

2.2 A Contrastive Probe

While the above study indicates that there may be structural similarities between the two spaces, Procrustes analysis makes strong assumptions about the underlying geometry of the two spaces, which may be overly restrictive. We now describe a contrastive probe that is more flexible than the Procrustes "probe", which will be trained on a set of held-in objects and and tested on how it generalizes to a set of held-out objects.

Our probe uses the following (learned) similarity function between the language and sound representations based on cosine similarity,

$$\mathrm{sim}(\mathrm{text}(c), \mathrm{sound}(c)) = \frac{\langle \mathbf{W}_1 f_{\mathrm{LM}}(\mathrm{text}(c)), \mathbf{W}_2 f_{\mathrm{AM}}(\mathrm{sound}\, c)\rangle}{\|\mathbf{W}_1 f_{\mathrm{LM}}(\mathrm{text}(c))\| \, \|\mathbf{W}_2 f_{\mathrm{AM}}(\mathrm{sound}\, c)\|},$$

with learnable linear transformations $\mathbf{W}_1 \in \mathbb{R}^{d \times d_1}, \mathbf{W}_2 \in \mathbb{R}^{d \times d_2}$ that project the text and audio embeddings into a common space. The above transformations can be learned in various ways; in this paper we use the standard contrastive loss objective,

$$L(\mathcal{C}) = \sum_{c \in \mathcal{C}} \Big(-\mathrm{sim}(\mathrm{text}(c), \mathrm{sound}(c))/\tau + \log \sum_{c' \in N(c)} \exp\left(\mathrm{sim}(\mathrm{text}(c), \mathrm{sound}(c'))/\tau\right) \Big),$$

where $N(c) \subseteq \mathcal{C} \setminus \{c\}$ is a set of randomly sampled negative samples and $\tau > 0$ is a temperature term.

Suppose we partition $\mathcal{C}$ into mutually exclusive train and test sets, $\mathcal{C} = \mathcal{C}^{\text{train}} \cup \mathcal{C}^{\text{test}}$. If we learn $\mathbf{W}_1, \mathbf{W}_2$ via minimizing $L(\mathcal{C}^{\text{train}})$ and these transformations generalize to $\mathcal{C}^{\text{test}}$, i.e., for $c \in \mathcal{C}^{\text{test}}$

$$c = \arg\max_{c' \in \mathcal{C}} \ \mathrm{sim}(\mathrm{text}(c'), \mathrm{sound}(c)),$$

then this would suggest that the language representation space $\mathcal{C}_{\text{language}}$ is structurally similar to the sound representation space $\mathcal{C}_{\text{sound}}$. (Note that the $\arg\max$ is over the entire set $\mathcal{C}$). We thus use accuracy@K over $\mathcal{C}^{\text{test}}$, where a prediction as counted as correct of the correct label is in the set of top K most similar objects, to evaluate the alignment between a language model f_{LM} and an audio model f_{AM}. See Figure 2 for an overview.

3 Experimental Setup

3.1 Models

Language representations. We test representations from a variety of text models, including word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), T-5 (Raffel et al., 2020), and LLaMA (Touvron et al., 2023). We also include several model versions within a family.

To extract the language representations from the above models for a given object c, we obtain a sentence using the template "`the sound of` c" and average the contextualized representations for the tokens corresponding to c within the resulting sentence.[6] Prior work on aligning language representations to visual representations have made use of natural sentences containing c (Ilharco et al., 2021; Li et al., 2023). In order to eliminate confounding factors that may arise from the other tokens in a sentence, we went with a simple templated approach for extracting the language representations.

Sound representations. We test audio embeddings from three models: AudioMAE (Huang et al., 2022), PaSST (Koutini et al., 2021) and PANN (Kong et al., 2020).

AudioMAE is a transformer-based model trained as a masked autoencoder on audio spectrograms. AudioMAE is pretrained via self-supervision on the AudioSet dataset (Gemmeke et al., 2017), which contains approximately 2 million segments of audio snippets from YouTube along with annotated labels that describe the sound event of the audio snippet (e.g., `dog`, `cat`, `aircraft`, ...).[7] Self-supervised AudioMAE is trained only on the spectrogram inputs of AudioSet. We also experiment with a supervised, finetuned version of AudioMAE (AudioMAE-FT) that is finetuned as an audio classification model on the same AudioSet dataset.

PaSST is also a spectrogram-based transformer model that has been trained as an audio classification model to predict sound events. PaSST is initialized from a vision transformer pretrained on ImageNet, which was shown to improve performance despite the difference in modalities (Gong et al., 2021).[8] PaSST performs two stages of supervised training: large-scale supervised learning on the broad AudioSet dataset, followed by smaller-scale finetuning on the FSD50K dataset (Fonseca et al., 2021), which is another (more freely-licensed) sound event classification dataset that inherits AudioSet's ontology.

Finally, PANN is a CNN-based model that is trained with supervision on the AudioSet dataset. Unlike the above models whose input is in the frequency domain (i.e., spectrograms), PANN operates directly over the time domain. We use the CNN14 version of PANN.

3.2 Dataset

Our main probing experiments are conducted on the FSD50K dataset (Fonseca et al., 2021), which includes around 50K audio clips with their annotated sound event classes with lengths ranging from 0.3-30 seconds. As some of the classes only have a few examples, we select the top 100 most frequent classes. In this case, each class has at least 117 audio samples.

Out of 100 classes, we randomly select 70 classes as the training object set $\mathcal{C}^{\text{train}}$, and learn the linear transformations $\mathbf{W}_1$ and $\mathbf{W}_2$ via the contrastive loss as described in §2.[9] We then apply the probe on the 30 held-out classes $\mathcal{C}^{\text{test}}$ to obtain an accuracy@K metric. For each audio snippet associated with c, prediction via retrieval is done over a superset of $\mathcal{C}$, in particular the set of 144 most frequent classes that were part of FSD50K (i.e., the retrieval set is over classes that were even outside the training set). This increases the difficulty of the task, and a similar approach was adopted in the context of aligning text and vision representations (Li et al., 2023). We repeat this training and testing over 5 different random partitions of $\mathcal{C}$ (each with a 70/30 split), and report the average accuracy@3 (over 144 classes) for audio snippets in the test set.

3.3 Hyperparameters

We performed a light grid search over the hyperparameters, in particular the learning rate $\alpha \in \{10^{-3}, 10^{-4}\}$, temperature coefficient $\tau \in \{0.07, 0.2\}$, number of negative samples $N(c) \in \{64, 128\}$. We use a batch size of 32 and train for 20 epochs. Importantly, we found the optimal hyperparameters (and the optimal epoch for early stopping) based on a *held-in* validation set that was randomly sampled from the training set. That is, none of the hyperparameters were tuned based on held-out performance on C^{test}.

3.4 Control Task

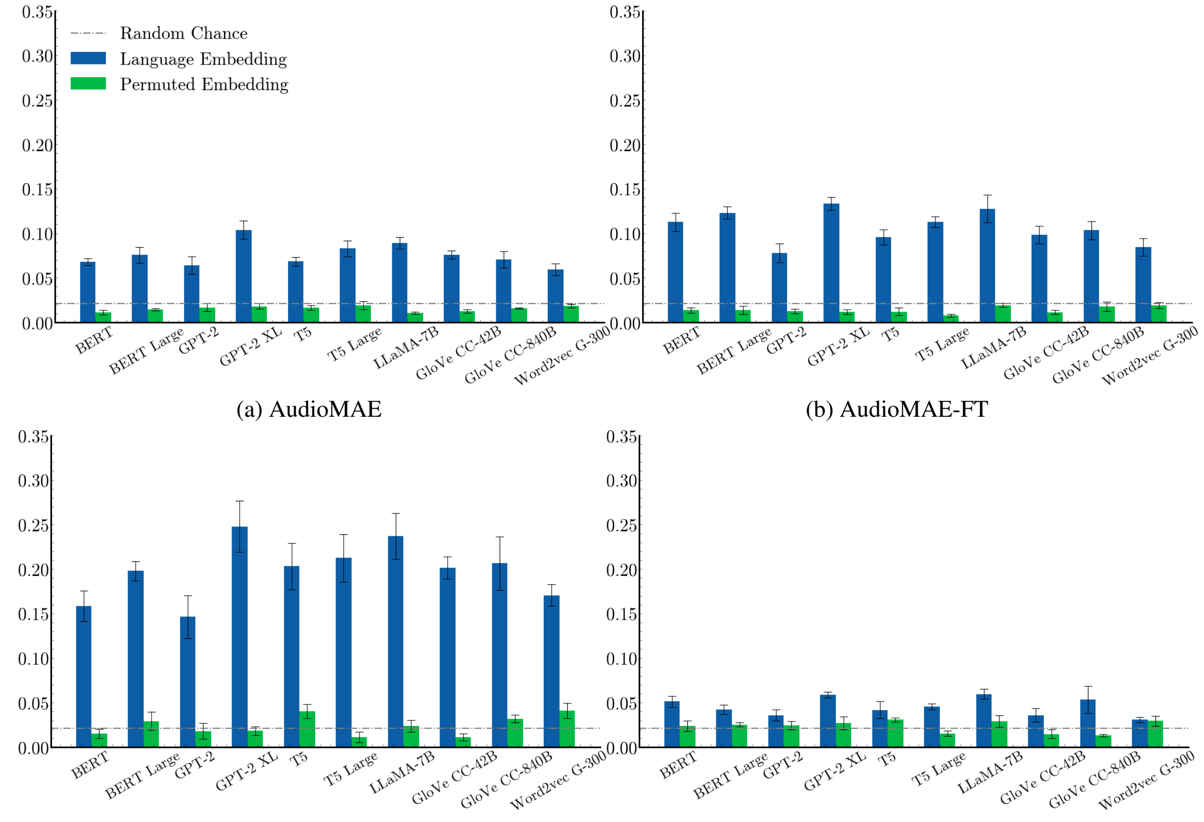
Because we evaluate accuracy only on the set of held-out objects, the usual caveats associated with probing (i.e., whether the above-chance performance is due to an LM's representations' meaningfully encoding the phenomena in question, or due to the probe's learning the task) are less of an

[6] For word2vec and GloVe, we just average the representation of "c".

[7] Since the AudioSet dataset contains examples of human speech, AudioMAE's sound representation is arguably not completely independent of language. The set of possible labels for human speech snippets in the dataset are {`male speech`, `female speech`, `child speech`, `conversation`, `narration`, `babbling`, `speech synthesizer`}.

[8] Thus PaSST representations potentially encode even more human perception-like priors.

[9] While the contrastive loss is presented assuming a single sample for $\text{sound}(c)$, in practice we train over multiple audio examples for a single object, instead of averaging the audio representations to obtain a single sound representation as was done in the Procrustes analysis (§2.1).



(a) AudioMAE

(b) AudioMAE-FT

(c) PaSST

(d) PANN

Figure 3: Accuracy@3 for the different language and sound representations. Green bars show the accuracy of the permuted embedding control task, where the text representations are randomly permuted. Error bars show standard error of the mean across 5 runs. Dotted line shows random chance performance, which is 2.08%.

issue. However, there may be other factors that may be contributing to above-chance performance, for example, the overall geometry of the respective representation spaces. We thus follow Hewitt and Liang (2019) and also compare the performance of our probes against a control task where we randomly permute the text embeddings.

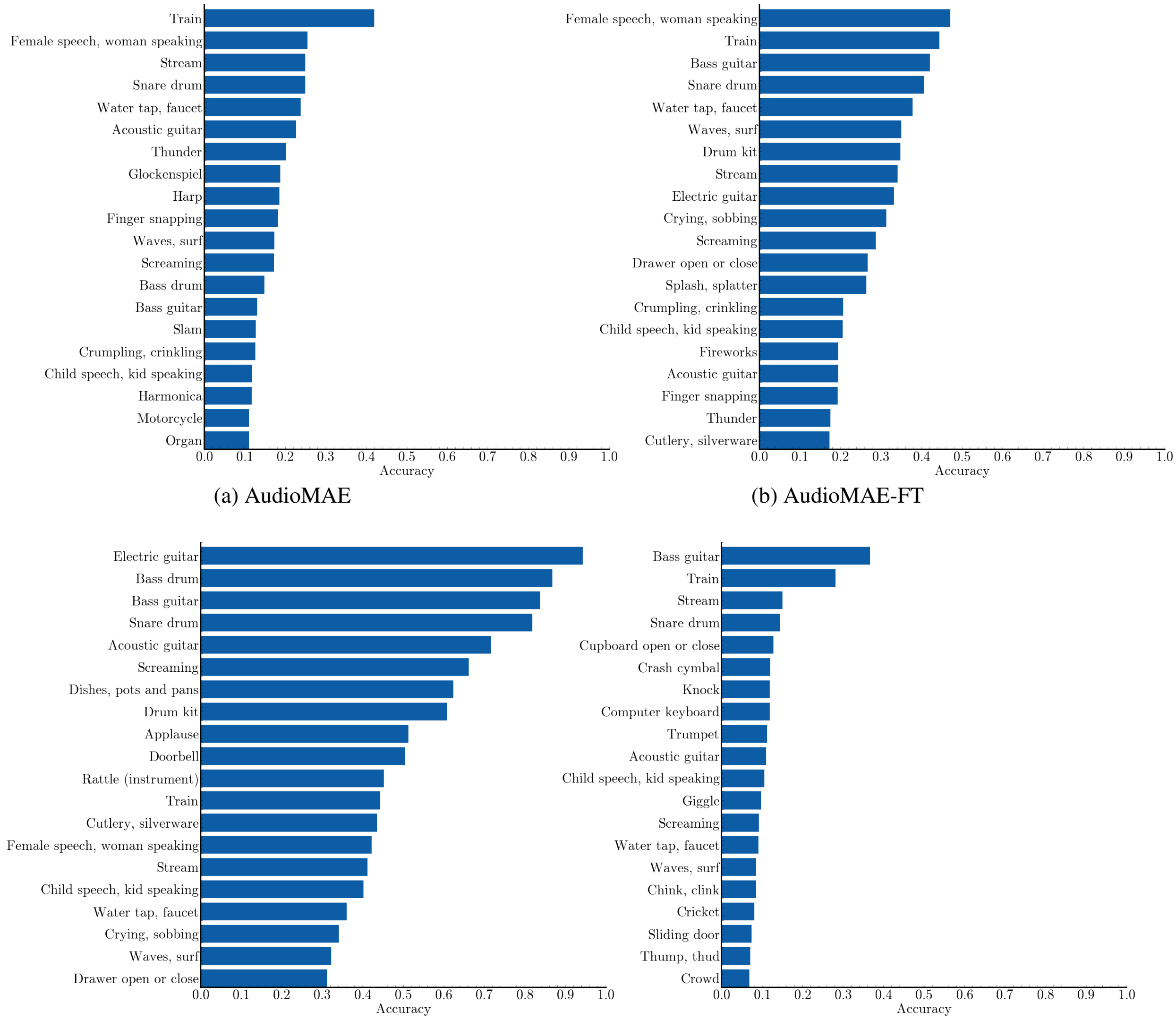
4 Results

Figure 3 shows the accuracy@3 metric for the different language/sound representation combinations. In the appendix, we report the full numeric values, including accuracy@1 and standard error across the 5 runs. We find that most language models perform well above chance. Moreover, within a model family, larger models almost always outperform their smaller siblings (e.g., BERT-Large vs. BERT, GPT-2-XL vs. GPT-2, T5-Large vs. T5). However, there is significant variation across families and larger models aren't always better across models (e.g., GPT-2-XL vs. LLaMA-7b). Despite their simplicity, word2vec and GloVe obtain nontrivial performance, sometimes outperforming much more sophisticated models.

Across the different audio models, we find that alignment is overall better for sound representations from PaSST, which arguably is most aligned to human perception insofar as it is pretrained as an image classifier, and then finetuned as a sound event classifier. The underperformance of AudioMAE (which is pretrained only on spectrograms via self-supervision and thus likely to focus only on acoustic information) against AudioMAE-FT (which is finetuned as a supervised classifier on top of AudioMAE and thus likely to additionally encode auditory—i.e., human perception-like—information) further highlights the importance of human-like representations that emerge from learning to predict human-derived labels. However, despite being trained as a supervised model, PANN performs the worst. This may be due to the fact that PANN's audio input is in the time domain, as well as the fact that PANN uses a CNN architecture instead of a transformer.

4.1 Analysis

What are the classes for which the probes generalize particularly well? In Figure 4 we show



(a) AudioMAE

(b) AudioMAE-FT

(c) PaSST

(d) PANN

Figure 4: Classes that had the best accuracies (as measured by accuracy@3) for the different sound representations. We measure the accuracies across all 5 train/test sets, and average across the different language models.

the classes that had the best accuracies (averaged across the language representations) across the five runs. We qualitatively find that classes that correspond to human speech, as well as instruments, seem to generalize well.

In Table 1 we perform a deeper analysis of GPT-2-XL, the best-performing language representation. For each class in Table 1 (ranked by accuracy), we show the top 3 closest classes in the set of training classes as measured by similarity in language/sound space.[10] In many cases, the top 3 retrieved classes are similar in both spaces, indicating structural similarities.

We next analyze whether classes that obtain high accuracies are in general similar across the different language models. For two language representations, we calculate Spearman's rank correlation between the accuracies of classes in the test set. We average this rank correlation across the 5 runs, which produces a measure of how similar the two language representations are in terms of their ability to encode sound information. Figure 5 shows the results for all pairwise correlations. As expected, language representations are generally the most similar within a language model family, although this is not always the case. Correlation is generally quite high across the different language representations despite the differences in model architectures, size, and training data; this potentially implies that there is a common set of classes for which sound representation is meaningfully encoded.

[10] Note that training accuracies were extremely high (e.g., >97%) for most classes. Therefore, it is not the case that (for example) "`Electric guitar`" accuracy is high because the text embedding for "`Electric guitar`" is the closest language embedding for *all* audio snippets.

			3 closest classes in language space			3 closest classes in sound space		
OOD Classes	**Acc@1**	**Acc@3**	**Class 1**	**Class 2**	**Class 3**	**Class 1**	**Class 2**	**Class 3**
AudioMAE								
Female spee..	0.75	0.93	Male speech	Yell	Whispering	Male speech	Yell	Whispering
Bass guitar	0.19	0.78	Acoustic gu..	Electric gu..	Snare drum	Acoustic gu..	Bowed strin..	Drum kit
Stream	0.09	0.64	Waves	Water tap	Sink	Water tap	Waves	Sink
Fireworks	0.04	0.55	Thunder	Gunshot	Keys jangli..	Gunshot	Thunder	Hammer
Burping	0.04	0.34	Chewing	Livestock	Fart	Chewing	Rattle	Livestock
Harmonica	0.02	0.23	Organ	Marimba	Ringtone	Wind instru..	Snare drum	Organ
AudioMAE-FT								
Female spee..	0.72	0.97	Male speech	Child speec..	Yell	Male speech	Yell	Screaming
Bass guitar	0.22	0.61	Electric gu..	Drum kit	Snare drum	Drum kit	Electric gu..	Acoustic gu..
Dishes	0.06	0.30	Cutlery	Hi-hat	Shatter	Cutlery	Hi-hat	Hiss
Bass drum	0.06	0.15	Snare drum	Drum kit	Hi-hat	Drum kit	Snare drum	Screaming
Wind	0.05	0.23	Waves	Whoosh	Thunder	Waves	Bark	Thunder
Giggle	0.04	0.72	Crying	Cough	Screaming	Crying	Cough	Screaming
PaSST								
Female spee..	0.71	0.98	Male speech	Yell	Child speec..	Male speech	Yell	Child speec..
Dishes	0.25	0.90	Cutlery	Keys jangli..	Sink	Cutlery	Keys jangli..	Rattle
Fixed-wing ..	0.09	0.64	Subway	Bus	Train	Subway	Gunshot	Bus
Scissors	0.05	0.45	Cutlery	Keys jangli..	Hammer	Cutlery	Keys jangli..	Rattle
Stream	0.03	0.34	Drip	Splash	Waves	Splash	Waves	Water tap
Bass drum	0.02	0.91	Drum kit	Snare drum	Acoustic gu..	Drum kit	Snare drum	Rattle
PANN								
Bass guitar	0.35	0.87	Acoustic gu..	Bowed strin..	Piano	Piano	Acoustic gu..	Bowed strin..
Stream	0.08	0.45	Waves	Train	Sink	Waves	Sawing	Buzz
Zipper	0.02	0.06	Writing	Cutlery	Camera	Crack	Shatter	Drip
Harp	0.00	0.00	Glockenspie..	Piano	Marimba	Organ	Chirp	Glockenspie..
Wind	0.00	0.00	Wind instru..	Buzz	Waves	Tick-tock	Sawing	Hiss
Walk	0.00	0.00	Knock	Run	Child speec..	Run	Crumpling	Hiss

Table 1: For the GPT-2-XL probe we we show the top 6 classes for which accuracy was the highest for each audio representation (for a given data split). For each class (which has multiple audio snippets associated with the class), we show the three closest classes as measured by cosine similarity both in language representation space and in sound representation space.

Audio Model	Linear	Non-Linear	Procrustes
AudioMAE	0.11	0.10	0.09
AudioMAE-FT	0.08	0.08	0.03
PaSST	0.20	0.16	0.17
PANN	0.05	0.04	0.02

Table 2: Generalization performance (accuracy@3) of different probes, where the performance is averaged across all language model representations.

4.2 Probing Method

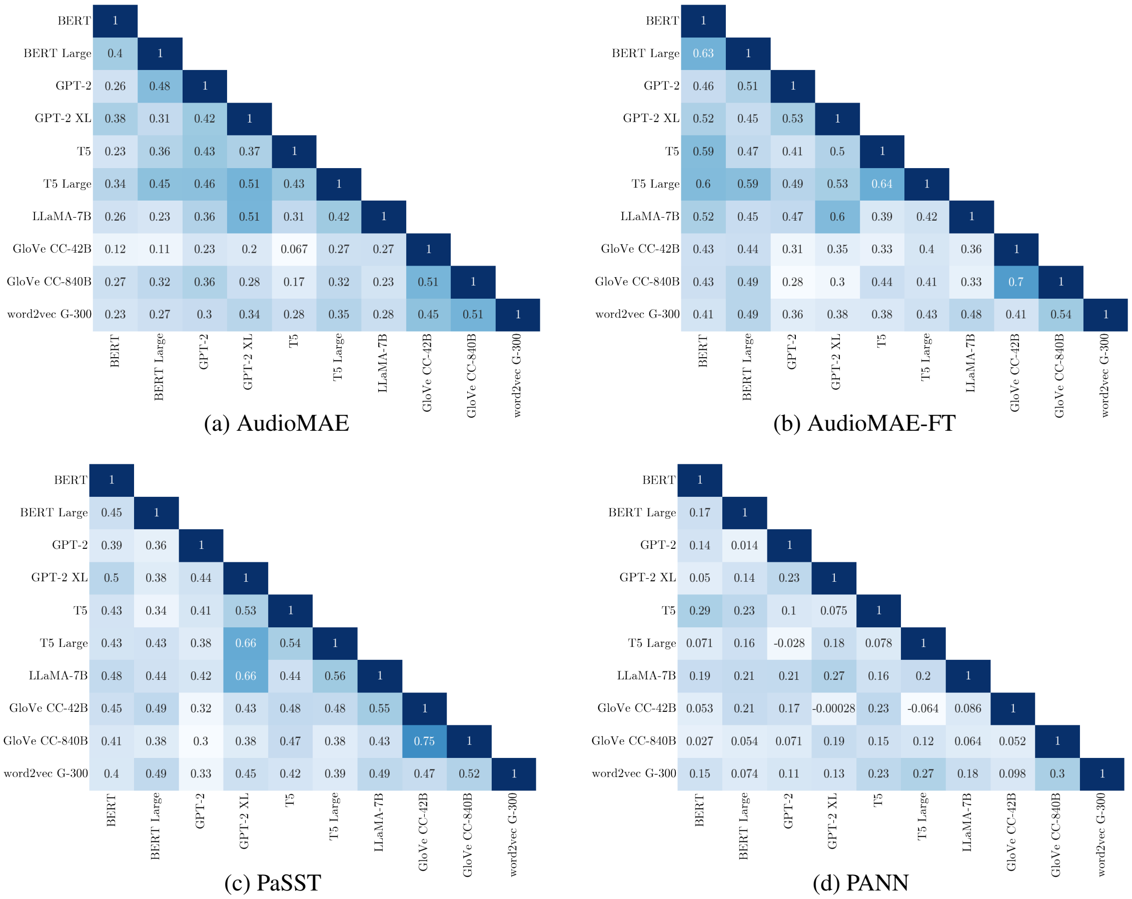
Our primary results make use of a contrastive loss with linear transformations applied to the language/sound representations. We additionally explore two other probes: the Procrustes probe discussed in §2.1 where we learn the matrix $\mathbf{Q}$ only based on the objects in $\mathcal{C}^{\text{train}}$; and a non-linear probe where we apply a ReLU non-linearity after projecting when calculating the similarity function.

The results are shown Table 2. We generally find that Procrustes probes, which minimize the MSE and additionally constrain the transformations to be orthogonal, underperform the contrastive loss probes. The linear probe outperforms the non-linear probe.

5 Discussion and Limitations

Our work, along with the line of work on aligning language model representations to grounded representations, provides evidence that modeling statistical correlations among surface-form text could lead to learning nontrivial structures about the real world. In hindsight, this is perhaps not so surprising; both language and sound are different "projections" of the same physical world, and thus it is not inconceivable that models trained on the respective modalities represent (some) aspects of the original physical world in a similar way.

More generally, the fact that current language models (and foundation models more generally) are trained only on "raw form" (such as word pieces, sound waves, pixels, etc.) is not an inherent limitation on their ability to learn physically grounded conceptual spaces. These models are typically trained (implicitly or explicitly) to compress their training data into their parameters; insofar as good compression can be achieved by learning the underlying generative process, it is possible that aspects of the physical world which were involved in the generation of language could be learned just through form-only training. Nonetheless, form-only training is likely to be highly data-inefficient.



(a) AudioMAE

(b) AudioMAE-FT

(c) PaSST

(d) PANN

Figure 5: Rank correlation of accuracies of classes within the test set between language representations, where the correlations are averaged across the five runs.

This work only explored whether sound representations that were learned by an auxiliary audio processing model were encoded through text. Here we found that sound representations that are more likely to encode auditory (i.e., human perception-like) information were more aligned to the text representations than sound representations from purely self-supervised models which were just trained on spectrograms, and thus more likely to encode acoustic information. However, even the self-supervised audio models implicitly encode human perception-aligned priors given that the input data consisted of snippets of audio that corresponded to different sound events, which itself is derived from humans. It would be interesting to see whether it is possible to probe out even more low-level representations of objects (e.g., raw spectrograms, pixels) from language models. Similarly, as discussed in footnote 7 our audio representations are not completely independent of language as their training sets included a significant amount of human speech. It would therefore be interesting to see if audio models trained without any human speech learn representatons that can be aligned to language models.

6 Related Work

Probing language models. Language models have been shown to encode much linguistic information in their contextualized representations (Tenney et al., 2019; Liu et al., 2019; Jawahar et al., 2019) and attention distributions (Clark et al., 2019; Vig and Belinkov, 2019). Building on top of these more linguistically-oriented probes, there has been mounting recent evidence that language models trained on just text are able to meaningfully encode a surprising amount of grounded or extralinguistic information, such as color (Abdou et al., 2021), direction (Patel and Pavlick, 2021), size (Zhang et al., 2020; Grand et al., 2022), geography (Konkol et al., 2017; Liétard et al., 2021; Faisal and Anastasopoulos, 2023; Chen et al., 2023), time (Gurnee and Tegmark, 2023), visual representations (Ilharco et al., 2021; Merullo et al., 2022; Li et al., 2023), character-level information of wordpieces (Kaushal and Mahowald, 2022), and representations of meaning (Li et al., 2021). Our work extends this line of to sounds and investigates the extent to which language models trained on text-only can encode auditory representations.

Our work is also related to the line of work investigating whether a model that has been trained on raw outputs of a synthetic environment can acquire "true" representations of that environment. Examples of such environments include Othello (Li et al., 2022), chess (Toshniwal et al., 2021), and toy grid worlds (Yun et al., 2023; Jin and Rinard, 2023).

Meaning in language models. Whether language models can acquire meaning and understanding from being trained on form alone is the subject of much debate (Bender and Koller, 2020; Merrill et al., 2021; Piantadosi and Hill, 2022; Pavlick, 2023; Søgaard, 2023). In operationalizations of meaning which do not rely on explicit reference to the external world, the fact that the geometry of language models' representation spaces is structurally related to the geometry of grounded representations could be construed as evidence for these models' acquiring meaning in some broad sense.

7 Conclusion

We probe text-only language models for whether their representations of an object contain grounded representations of the sounds of the same object. We find that this is indeed the case, and a contrastive probe can often generalize zero-shot to object classes not seen during training.

Acknowledgments

This study was partially supported by funds from the MIT-IBM Watson AI Lab.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yida Chen, Yixian Gan, Sijia Li, Li Yao, and Xiaohan Zhao. 2023. More than correlation: Do large language models learn causal representations of space? *arXiv preprint arXiv:2312.16257*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2023. Geographic and geopolitical biases of language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163, Singapore. Association for Computational Linguistics.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. In *In Proceedings of NeurIPS*.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2021. Probing Contextual Language Models for Common Ground with Visual Representations. ArXiv:2005.00619 [cs].

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Charles Jin and Martin Rinard. 2023. Evidence of meaning in language models trained on programs. *arXiv preprint arXiv:2305.11169*.

Ayush Kaushal and Kyle Mahowald. 2022. What do tokens know about their characters and how do they know it? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2487–2507, Seattle, United States. Association for Computational Linguistics.

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

Michal Konkol, Tomáš Brychcín, Michal Nykl, and Tomáš Hercig. 2017. Geographical evaluation of word embeddings. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 224–232, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. 2021. Efficient training of audio transformers with patchout. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022-September:2753–2757.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Jiaang Li, Yova Kementchedjhieva, and Anders Søgaard. 2023. Implications of the Convergence of Language and Vision Model Geometries. ArXiv:2302.06555 [cs].

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Bastien Liétard, Mostafa Abdou, and Anders Søgaard. 2021. Do language models know the way to rome? *arXiv preprint arXiv:2109.07971*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Anders Søgaard. 2023. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33(1):33–54.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. Learning chess blindfolded: Evaluating language models on state tracking. *arXiv preprint arXiv:2102.13249*, 2.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Tian Yun, Zilai Zeng, Kunal Handa, Ashish V Thapliyal, Bo Pang, Ellie Pavlick, and Chen Sun. 2023. Emergence of abstract state representations in embodied sequence modeling. *arXiv preprint arXiv:2311.02171*.

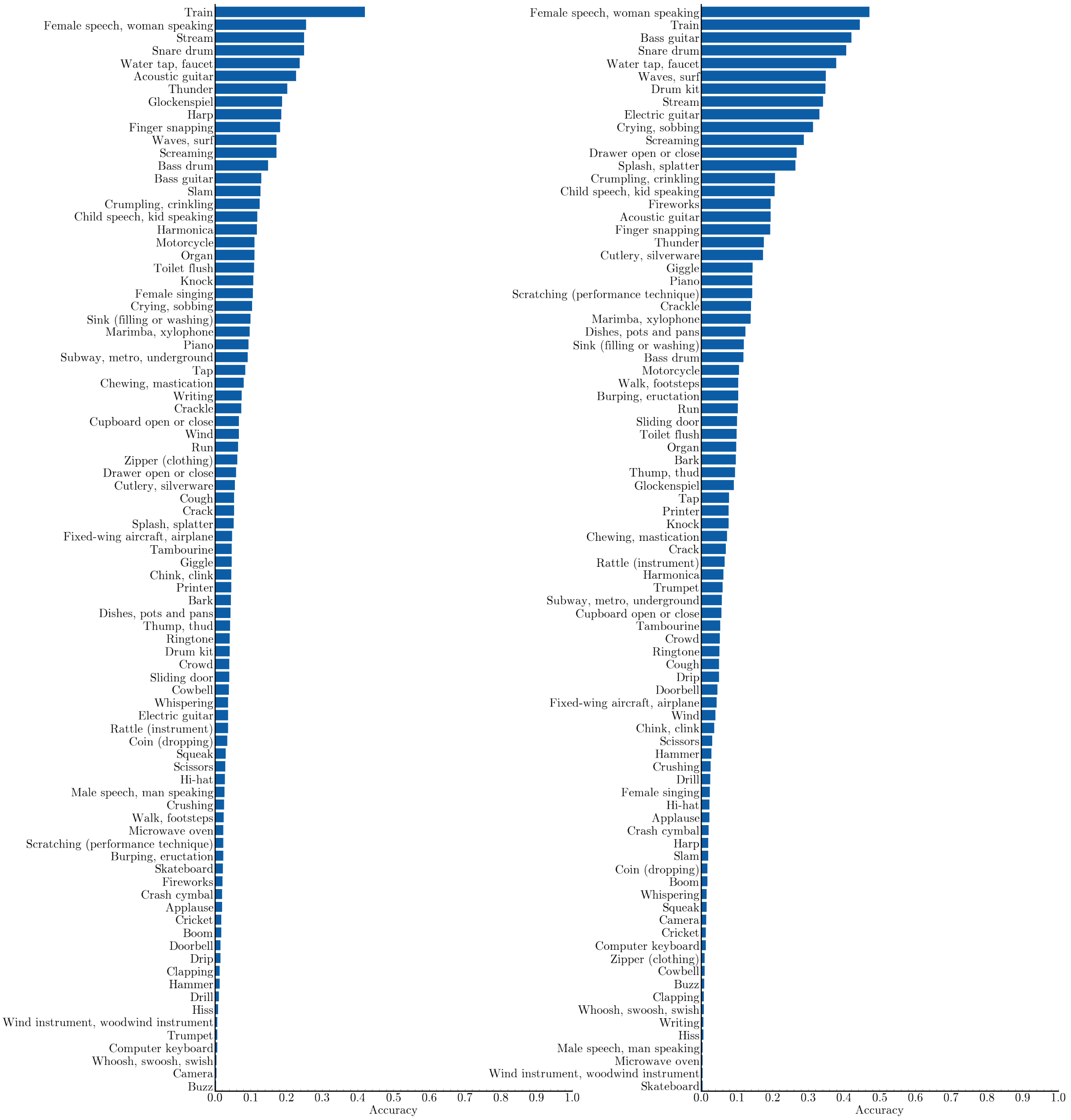
Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

A Appendix

We show the full numeric results for our main probing experiments in table 3 and for classes accuracy in figure 6.

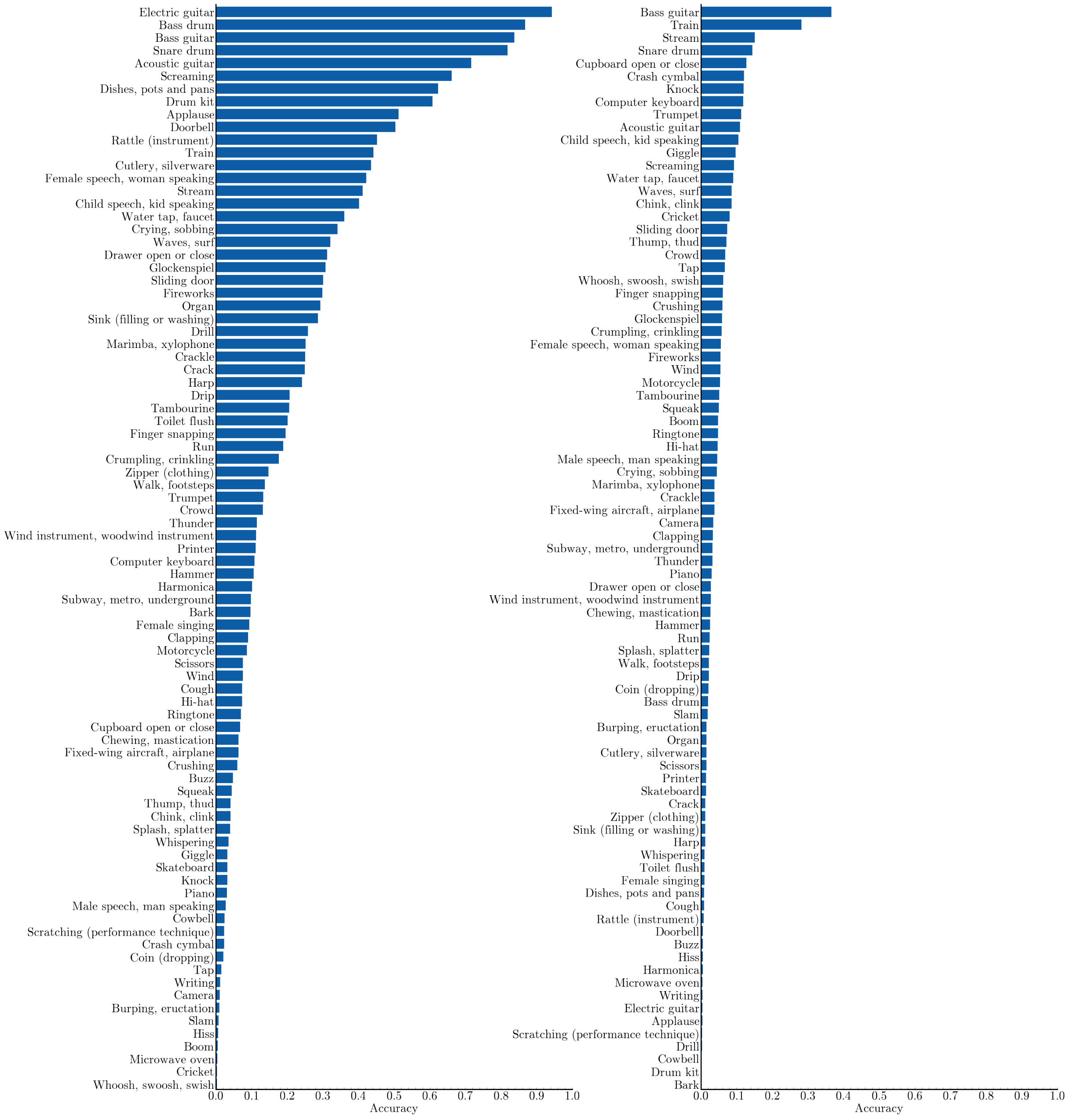
	Language Embedding		**Permuted Embedding**		**Random Init**	
Models	**A@1**	**A@3**	**A@1**	**A@3**	**A@1**	**A@3**
AudioMAE						
BERT	0.02 ± 0.0	0.07 ± 0.0	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.01 ± 0.01
BERT Large	0.02 ± 0.0	0.08 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.02 ± 0.01
GPT-2	0.01 ± 0.0	0.06 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.0 ± 0.0	0.02 ± 0.01
GPT-2 XL	0.02 ± 0.0	0.1 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.01 ± 0.0	0.03 ± 0.01
T5	0.02 ± 0.0	0.07 ± 0.0	0.0 ± 0.0	0.02 ± 0.0	0.02 ± 0.0	0.03 ± 0.01
T5 Large	0.02 ± 0.0	0.08 ± 0.01	0.01 ± 0.0	0.02 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
LLaMA-7B	0.02 ± 0.0	0.09 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.0	0.03 ± 0.0
GloVe CC-42B	0.02 ± 0.0	0.08 ± 0.0	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.0	0.03 ± 0.01
GloVe CC-840B	0.02 ± 0.0	0.07 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
word2vec GNews-300	0.01 ± 0.0	0.06 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
AudioMAE-FT						
BERT	0.02 ± 0.0	0.11 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.02 ± 0.0
BERT Large	0.02 ± 0.01	0.12 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.02 ± 0.0
GPT-2	0.01 ± 0.0	0.08 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.0	0.02 ± 0.0
GPT-2 XL	0.02 ± 0.0	0.13 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.01	0.03 ± 0.01
T5	0.01 ± 0.0	0.1 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
T5 Large	0.02 ± 0.0	0.11 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.01 ± 0.0
LLaMA-7B	0.02 ± 0.0	0.13 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.0 ± 0.0	0.01 ± 0.01
GloVe CC-42B	0.02 ± 0.0	0.1 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.02 ± 0.0
GloVe CC-840B	0.02 ± 0.0	0.1 ± 0.01	0.01 ± 0.0	0.02 ± 0.0	0.01 ± 0.0	0.03 ± 0.01
word2vec GNews-300	0.01 ± 0.0	0.08 ± 0.01	0.0 ± 0.0	0.02 ± 0.0	0.01 ± 0.0	0.01 ± 0.0
PaSST						
BERT	0.02 ± 0.01	0.16 ± 0.02	0.0 ± 0.0	0.02 ± 0.0	0.01 ± 0.01	0.02 ± 0.01
BERT Large	0.03 ± 0.01	0.2 ± 0.01	0.0 ± 0.0	0.03 ± 0.01	0.0 ± 0.0	0.02 ± 0.01
GPT-2	0.02 ± 0.0	0.15 ± 0.02	0.0 ± 0.0	0.02 ± 0.01	0.01 ± 0.01	0.03 ± 0.01
GPT-2 XL	0.04 ± 0.01	0.25 ± 0.03	0.0 ± 0.0	0.02 ± 0.0	0.01 ± 0.01	0.01 ± 0.01
T5	0.02 ± 0.01	0.2 ± 0.02	0.01 ± 0.0	0.04 ± 0.01	0.01 ± 0.01	0.03 ± 0.01
T5 Large	0.02 ± 0.01	0.21 ± 0.02	0.0 ± 0.0	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
LLaMA-7B	0.04 ± 0.01	0.24 ± 0.02	0.0 ± 0.0	0.02 ± 0.01	0.0 ± 0.0	0.01 ± 0.01
GloVe CC-42B	0.02 ± 0.01	0.2 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.01	0.01 ± 0.01
GloVe CC-840B	0.03 ± 0.01	0.21 ± 0.03	0.01 ± 0.0	0.03 ± 0.0	0.01 ± 0.01	0.01 ± 0.01
word2vec GNews-300	0.02 ± 0.0	0.17 ± 0.01	0.0 ± 0.0	0.04 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
PANN						
BERT	0.02 ± 0.0	0.05 ± 0.01	0.01 ± 0.0	0.02 ± 0.01	0.0 ± 0.0	0.02 ± 0.01
BERT Large	0.01 ± 0.0	0.04 ± 0.0	0.01 ± 0.0	0.03 ± 0.0	0.0 ± 0.0	0.01 ± 0.01
GPT-2	0.02 ± 0.01	0.04 ± 0.01	0.01 ± 0.0	0.02 ± 0.0	0.01 ± 0.01	0.03 ± 0.0
GPT-2 XL	0.02 ± 0.0	0.06 ± 0.0	0.01 ± 0.0	0.03 ± 0.01	0.0 ± 0.0	0.01 ± 0.01
T5	0.01 ± 0.0	0.04 ± 0.01	0.01 ± 0.0	0.03 ± 0.0	0.01 ± 0.01	0.03 ± 0.01
T5 Large	0.02 ± 0.0	0.05 ± 0.0	0.0 ± 0.0	0.02 ± 0.0	0.01 ± 0.01	0.02 ± 0.0
LLaMA-7B	0.02 ± 0.0	0.06 ± 0.01	0.01 ± 0.0	0.03 ± 0.01	0.0 ± 0.0	0.03 ± 0.01
GloVe CC-42B	0.01 ± 0.0	0.04 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.0 ± 0.0	0.01 ± 0.01
GloVe CC-840B	0.02 ± 0.01	0.05 ± 0.01	0.0 ± 0.0	0.01 ± 0.0	0.01 ± 0.01	0.01 ± 0.01
word2vec GNews-300	0.01 ± 0.0	0.03 ± 0.0	0.01 ± 0.0	0.03 ± 0.01	0.01 ± 0.01	0.01 ± 0.01

Table 3: Numeric values for accuracy@1 (A@1) and accuracy@3 (A@3) for our main sound probing experiments. We also show standard error of the mean across 5 runs. Random init refers to a probe trained over randomly initialized language/audio models.



(a) AudioMAE

(b) AudioMAE-FT



(c) PaSST

(d) PANN

Figure 6: All class accuracies for different audio models. Each class's accuracy is averaged across 5 train/test sets and different language models.