ENCODER-FREE MULTIMODAL LLMS LEARN IMPLICIT ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We empirically analyze two recently proposed encoder-free multimodal LLMs, Gemma-4 12B and Inkling-Small. Unlike previous multimodal LLMs, these models do not use modality-specific encoders to extract tokenized embeddings representing audio and image input; instead, they directly feed windows of samples, pixels, or spectral features to the Transformer language model. We analyze whether the early layers of these models learn to function similarly to dedicated audio and visual encoders. We find that these early layers do in fact act as "implicit encoders," encoding audio and visual inputs in a way that is largely invariant to their cross-modal context. Furthermore, when performing multimodal reasoning tasks, we show that independently encoding the modalities throughout these early model layers achieves the same accuracy as jointly encoding them from the first layer onwards. Finally, we use layer-wise probing to analyze the salient information captured by the implicit encoders and find their behavior to be similar to several popular encoder models.

1 INTRODUCTION

Multimodal large language models (LLMs) such as ChatGPT (OpenAI, 2023), Claude (Anthropic, 2024), Gemini (Gemini Team, Google, 2023), Qwen-Omni (Xu et al., 2025) natively process text alongside general audio and visual inputs. Because this style of model is capable of handling a wide range of both multimodal and single-modality tasks, they are rapidly displacing modality- and task-specific models. Most multimodal LLMs follow the same architectural blueprint: pre-trained, modality-specific encoders are used to extract tokenized representations of visual and audio inputs. A text LLM backbone is pre-trained separately, and the two are stitched together via a connector layer. The model is then adapted through connector-only, LoRA, or full fine-tuning.

However, as LLM backbones become larger and larger, some researchers have started to investigate whether the role of the encoders can be completely subsumed by the LLM itself. To this end, there have been several very recently introduced "encoder-free" multimodal LLMs which instead tokenize audio and visual inputs by applying sliding windows (possibly with additional lightweight processing, such as a linear projection) to low-level representations, such as image pixels, waveform samples, or spectral time-frequency energies. Recent examples of this style of model are Google's Gemma-4 12B (Gemma Team, Google DeepMind, 2026) and the Inkling model (Thinking Machines Lab, 2026) from Thinking Machines. The fact that these models still achieve strong performance on a range of tasks calls into question the necessity of separate audio and visual encoders. However, the pros and cons of this approach are still not fully understood. Encoder-free models offer a simpler, monolithic model architecture and have the potential to perform very early fusion of cross-modal information. However, current models (such as Gemma-4 12B) require their inputs to adhere to a specific multimodal format, and they may be less efficient to train than more modular models, especially if a large portion of their training time is spent re-learning representations that are highly similar to what a separate, pre-trained encoder would have learned anyway.

In this paper, we put forth the hypothesis that the encoder-free Gemma-4 12B and Inkling-Small models in fact *rediscover* audio and image encoders using their early Transformer layers. We empirically test for implicit encoder-like behavior in these models using several techniques. We first confirm the invariance of audio and image token representations across model layers as their cross-modal context is altered, and use activation patching (Meng et al., 2022; Zhang & Nanda, 2024) to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

show that independently encoding the modalities throughout the Transformer layers achieves similar performance, on a multimodal reasoning task, to jointly encoding them from the first layer onwards. Next, we perform layer-wise analysis of the speech audio representations learned by the "implicit encoder" layers of the LLMs, and find that they are highly similar to the representations of popular encoders such as WavLM (Chen et al., 2022) and HuBERT (Hsu et al., 2021). Finally, we test the "implicit encoder" layers *as encoders* by evaluating them on standard encoder evaluation benchmarks such as SUPERB (Yang et al., 2021) and MIEB (Xiao et al., 2025), finding that on some tasks they are competitive, but on other tasks they fall short of existing encoders that have far smaller parameter counts. The outcomes of all of these experiments support our hypothesis that Gemma-4 12B and Inkling-Small do learn audio and image encoder-like behavior within their early layers.

2 Related Work

2.1 Pre-Trained Encoders and Multimodal LLMs

The dominant paradigm for building multimodal LLMs is to attach one or more pre-trained, modality-specific encoders to a pre-trained text LLM. Vision Transformers (ViTs) (Dosovitskiy et al., 2021) such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023) are often used as image encoders, while pre-trained speech recognizers such as Whisper (Radford et al., 2023) and USM (Zhang et al., 2023) or self-supervised models such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) are often used to encode audio. The encoder output is then mapped into the backbone LLM's token embedding space by a "connector", which is often a linear projection or lightweight MLP as in LLaVA (Liu et al., 2023), SALMONN (Tang et al., 2024), and Qwen2-Audio Chu et al. (2024). Multimodal models such as Qwen-Omni Xu et al. (2025), MOLMo Deitke et al. (2024), and Gemini Gemini Team, Google (2023) extend this framework to three or more modalities by using parallel encoders for audio and visual inputs, alongside text tokenizers typically based on byte-pair encoding (Sennrich et al., 2016) and a text token embedding table.

2.2 Encoder-Free Multimodal LLMs

A recently introduced type of multimodal LLM is the "encoder-free" architecture, in which raw pixel patches or windows of low-level audio features (waveform samples or spectral features) are projected into the input token space of a text LLM. The major motivations for this approach are twofold: first, it offers a simpler, monolithic model rather than an ensemble of pre-trained encoders assembled to a backbone LLM. Second, it allows the encoders to be jointly optimized with the LLM backbone, and potentially allows for richer fusion of cross-modal information from the earliest layers of the model. Early examples of this paradigm applied to Vision-and-Language models (VLMs) are Fuyu-8B (Bavishi et al., 2023), EVE (Diao et al., 2024; 2025), and SOLO (Chen et al., 2024). Gemma-4 12B (Gemma Team, Google DeepMind, 2026) and Inkling (Thinking Machines Lab, 2026) are the first examples of large-scale, open-weight, encoder-free LLMs that simultaneously handle text, audio, and image input. They form the basis of our studies in this paper, and so we discuss their architectures in detail in the following two paragraphs.

Gemma-4 12B is an open-weight, 48 layer, decoder-only Transformer LLM. It is capable of processing text, image, video, and audio inputs with a unique encoder-free architecture. While the other models in the Gemma-4 family utilize large, pre-trained ViT (Dosovitskiy et al., 2021) vision encoders and USM (Zhang et al., 2023) audio encoders, the Gemma-4 12B model was trained from scratch using only linear projections to embed image patches and waveform windows into its token embedding space. Audio waveform samples at a 16kHz sampling rate are captured with non-overlapping sliding windows of 640 samples (40ms) before being projected into the 3840-dimensional input token space with a single linear transform. Similarly, non-overlapping RGB image pixel patches of size 48x48x3 are projected into the same 3840-dimensional token embedding space before being concatenated with the text and audio inputs to the model. Finally, an important detail about Gemma-4 12B that pertains to our analysis is the fact that it uses interleaved global and sliding window attention across its Transformer layers at a 5:1 ratio. For all of our experiments, we use the instruction-tuned version of Gemma-4 12B.

Inkling is a recently-released, open-weight, multimodal LLM capable of processing text, image, and audio inputs. We conduct our experiments on the Inkling-Small model, a 42-layer, decoder-only, Tranformer LLM that uses a sparse Mixture of Experts (Shazeer et al., 2017) architecture. The model has 276B total and 12B active parameters, and was trained on 45T total tokens spanning text, audio, image, and video. Similarly to Gemma-4 12B, Inkling-Small interleaves global and sliding-window self-attention with a 5:1 ratio, and is also encoder-free, although its architecture differs from that of Gemma-4 12B. 40x40x3 patches of RGB image pixels are encoded with a 4-layer MLP into 4096-dimensional tokens. Audio frames are represented as dMel (Bai et al., 2024) quantized spectral features, which are mapped into the 4096-dimensional tokens via a linear projection.

2.3 Representation Analysis and Emergent Modularity

Many prior works have studied the representations that emerge within deep neural network models. A recurring theme throughout this work is the emergence of a depth-ordered hierarchy of abstraction. For example, Zeiler & Fergus (2014) showed that successive convolutional layers progress from representations of oriented edges through textures and object parts to whole-object detectors, and Pasad et al. (2021; 2023; 2024) made similar observations about how self-supervised speech audio models capture low-level acoustic features to phones and words with increasing layer depth. In the text domain, The "BERTology" literature (Rogers et al., 2020; Tenney et al., 2019) found that the BERT (Devlin et al., 2018) model "rediscovers" part of-speech tagging, parsing, named-entity recognition, semantic roles, and co-reference resolution at different layers in the model.

Recent work has carried these tools to LLMs and to multimodal LLMs. Lad et al. (2025) propose that LLMs share the same general stages of inference, while Wendler et al. (2024) finds that multilingual models pivot through an English-like concept space in their middle layers. Causal methods such as activation patching (Zhang & Nanda, 2024) test whether a representation is actually used downstream. Applied to multimodal LLMs, these methods have shown that visual token representations become progressively more language-aligned with depth (Neo et al., 2024), and that cross-modal integration concentrates in particular layers and token positions (Zhang et al., 2024). Our work extends some of these analysis techniques to encoder-free, multimodal LLMs and aims to uncover whether parts of these models exhibit emergent specialization that fulfills the same role as the modality encoders traditionally employed in multimodal LLMs.

3 Analysis of Gemma-4 12B and Inkling-Small

3.1 Context invariance of embeddings

Because Gemma-4 12B and Inkling-Small take as input a concatenated sequence of text, image, and audio tokens, these inputs have the opportunity to perform cross-modal contextualization starting at the very first layer of the model. If this contextualization does occur, then the same audio input encoded multiple times with different text or image contexts will result in different representations of the audio tokens. However, our hypothesis is that the early layers of these models are acting as modality-specific encoders, and therefore will not contextualize across modalities until a sufficiently abstract representation is computed. If this hypothesis is true, then the implication is that there exists some layer L such that the representation of an audio (or visual) input to the model will be invariant to changes in its cross-modal context up until layer L, after which it will start to diverge with varying context. We experimentally investigated the degree to which Gemma-4 12B and Inkling-Small obey this context invariance, using either audio or visual input as the anchor modality.

3.1.1 Context Invariance of Audio Inputs

We randomly sample 1,000 utterances from the dev-clean split of Librispeech (Panayotov et al., 2015), and form a neutral context for each of these by pairing each waveform with a neutral text string (unrelated to the speech waveform; the same string is used for all waveforms) and a black-to-white gradient image. We then encode each waveform jointly with the text and image inputs using both models, and extract their layer-wise outputs (3840-dimensional outputs across 48 layers for Gemma-4 12B, and 4096-dimensional outputs across 42 layers for Inkling-Small). These embeddings serve as the neutral reference representations.

Next, we replace the gradient image and/or neutral text with different inputs: **Transcription Prompt**: A text prompt instructing the model to transcribe the content of the input speech waveform into a word sequence. **Emotion prompt**: A text prompt instructing the model to recognize the emotional content of the input speech waveform. **Real photo**: A natural photograph, sampled from the MSCOCO (Lin et al., 2014) dataset. We re-encode each speech waveform 4 times with each of the models: once for each of the three context changes listed above, and a 4th time that simultaneously replaces the gradient image with a natural photo and the neutral text with the transcription prompt. Next, we use Centered Kernel Analysis (CKA) (Kornblith et al., 2019) to compute the layer-to-layer similarity between the neutral reference representations of the audio input tokens, vs. the same audio inputs with varying text and image contexts. The results of this analysis are depicted in Figure 1.

For Gemma-4 12B (Figure 1), across all of the context changes the CKA similarity of the audio token representations to their neutral reference remains high (well above 0.9) until layer 12, after which they gradually drop with increasing depth into the model. This suggests that the representations of the audio embeddings computed by the first 12 layers of Gemma-4 12B do not significantly change as a function of varying their textual and visual context. We would observe similar behavior if the model instead used a completely separate audio encoding trunk unable to attend to the visual or text inputs. The bulk of the drop in CKA similarity to the reference embeddings happens gradually between layers 12 and 48, indicating that cross-modal contextualization takes place throughout the latter 75% of the model layers, presumably only *after* the audio token representations have reached a sufficiently high level of abstraction. We analyze this in more depth in Section 3.5.

Inkling-Small (Figure 1) shows an even more striking trend: aside from a small dip at layer 14 and a significant drop at the very last layer, the CKA similarity between the altered-context audio representations and the neutral-context representations remains very high across all layers of the model. This suggests that Inkling-Small does not contextualize the audio token representations with cross-modal information to a significant degree until the very last layer of the model. We analyze exactly where Inkling-Small performs cross-modal contextualization later in Section 3.2.

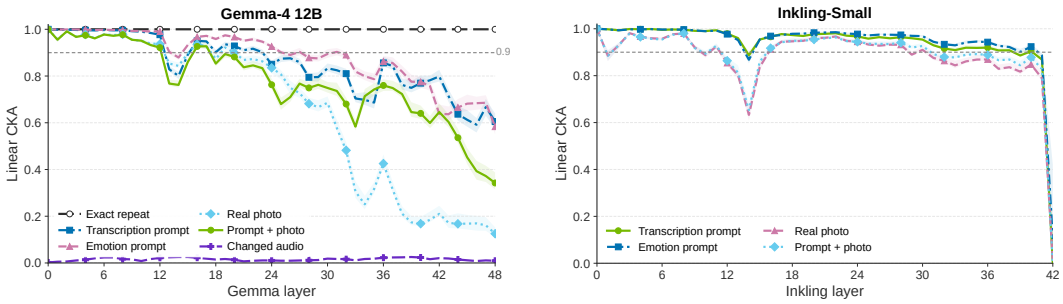


Figure 1: Layer-wise CKA similarity of audio tokens encoded under varying image and text contexts vs. audio tokens encoded with a neutral context.

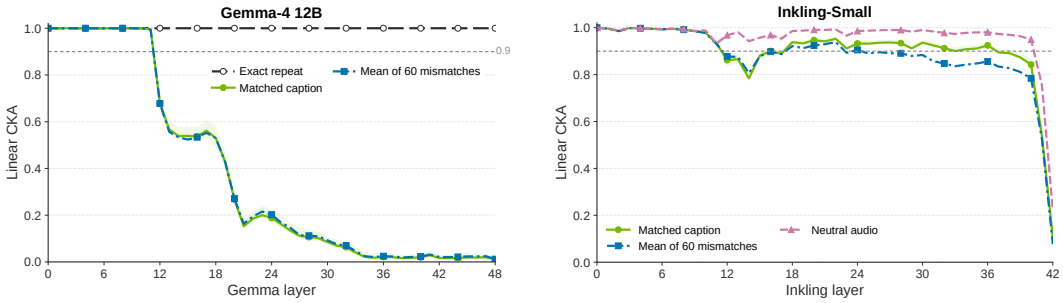
3.1.2 Context Invariance of Visual Inputs

Next, we conduct a similar experiment but use the visual inputs as the anchor modality. Gemma-4 12B's canonical token layout places text tokens first, followed by image tokens, and finally followed by audio tokens. Because the model uses causally-masked self-attention, the image tokens do not have an opportunity to attend to the audio tokens. We therefore restrict our analysis of the context invariance of image embeddings to comparisons between different textual contexts. First, we randomly sample 797 images from MSCOCO (Lin et al., 2014), and pair each one with three different text contexts: **No caption**: Blank whitespace is provided to the model instead of a caption. **Matched caption**: A ground-truth matching caption for each image from MSCOCO is sampled and provided as input to the model along with its paired image. **Mismatched caption**: A randomly sampled caption belonging to a different MSCOCO image is used as the text context.

We jointly encode each image with the "No caption" contextual condition and extract the per-layer outputs from Gemma-4 12B to form the reference embeddings. Next, we encode each image with its matched caption. Finally, we encode each image 60 times with 60 different mismatched cap-

tions, and average the embeddings. We then compute the CKA similarity between the reference embeddings and the embeddings computed in the context of a matched caption, and finally repeat this procedure between the reference embeddings and mismatched caption embeddings. The results of this experiment are depicted in Figure 2.

In the case of Gemma-4 12B (Figure 2(a)), we again observe similar curves to what we saw when analyzing the encoding of audio tokens with the same model (Figure 1): the CKA similarity stays extremely high during the early layers of the model, staying close to 1.0 until layer 12, at which point it rapidly drops. This provides additional evidence that the early layers of Gemma-4 12B are encoding audio and visual inputs into higher-level representations *before* allowing for cross-modal contextualization, which only begins at the middle layers of the model. Inkling-Small (Figure 2(b)) again behaves differently from Gemma-4 12B, with its image patch embeddings staying almost completely invariant to cross-modal context until layer 10, at which point they become slightly sensitive to context and remain so until layer 42. This behavior is very similar to what we observed when examining how the same model encodes audio tokens (Figure 1).



(a) Analysis of image context sensitivity for the Gemma-4 12B Model.

(b) Analysis of image context sensitivity for the Inkling-Small Model.

Figure 2: Layer-wise CKA similarity of image tokens encoded under varying text and audio contexts vs. image tokens encoded with a neutral context.

3.2 Independent vs. Joint Encoding of Modalities for Cross-Modal Reasoning

Our next experiment follows up on our context-invariance analysis by tasking the models with a question answering task that requires simultaneous reasoning across text, image, and audio inputs. Nominally, both Gemma-4 12B and Inkling-Small would take all three modalities as input simultaneously and perform joint encoding. In addition to this condition, we perform a second forward pass that uses a form of activation patching (Zhang & Nanda, 2024): We separately encode the tokens belonging to each modality (isolated from all other modalities, with one forward pass per modality) up until layer L, extract these neutral-context representations at layer L for all modalities, concatenate them together according to the model's expected token layout, and then feed these concatenated tokens into the remainder of the model beginning at layer $L + 1$. Informed by our analysis in Section 3.1, we hypothesize that if the inputs to the models are invariant to changes in their cross-modal context up until some layer L, then there should be no difference in their performance on the question answering task using jointly vs. independently encoded inputs up until layer L; after layer L, we should expect to see changes in the performance of the model.

To implement this experiment, we use the OmniBench (Li et al., 2024) dataset which is comprised of 1,142 quadruplets, each containing a text question, a visual image, an audio waveform, and a text answer. The evaluation examples in OmniBench were constructed in such a way that each question should only be answerable by taking into account both the image and the audio waveform. For each of the 1,142 examples in OmniBench, we separately encode the question text, image, and audio waveform with three separate forward passes through the model. When encoding one of these modalities, we preserve its positional information by filling in the tokens belonging to the other modalities with padding. This results in the token embeddings retaining their positional embeddings, and also keeps the number of tokens in their surrounding context constant.

The results of our OmniBench analysis on both models is shown in Figure 3. The blue dashed line indicates the 43.8% baseline accuracy that Inkling-Small achieves on the task when all modalities

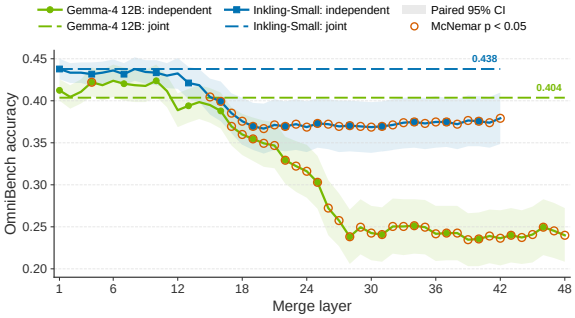


Figure 3: Performance of Gemma-4 12B and Inkling-Small on OmniBench as a function of the model layer where independently-encoded modality representations are merged.

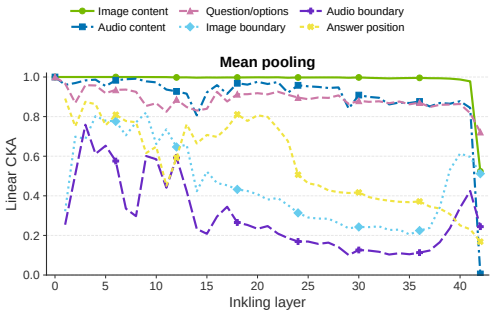


Figure 4: Similarity of jointly versus independently encoded representations for different token segments: content tokens (image, audio, text) and boundary tokens.

Table 1: Performance of the first 12 layers of Gemma-4 12B and Inkling-Small compared to baseline SSL models on SUPERB. We also report the performance of our distilled Gemma models (Section 3.6)

Model	Mean-Pooled Embeddings				Frame-wise Embeddings						
	KS	IC	SID	ER	PR	ASR	QbE	SF		ASV	SD
	Acc↑	Acc↑	Acc↑	Acc↑	PER↓	WER↓	MTWV↑	F1↑	CER↓	EER↓	DER↓
FBANK	8.63	9.10	8.5E-4	35.39	82.01	23.18	0.0058	69.64	52.94	9.56	10.05
wav2vec 2.0 Base	96.23	92.35	75.18	63.43	5.74	6.43	0.023	88.30	24.77	6.02	6.08
wav2vec 2.0 Large	96.66	95.28	86.14	65.64	4.75	3.75	0.049	87.11	27.31	5.65	5.62
HuBERT Base	96.30	98.34	81.42	64.92	5.41	6.42	0.074	88.53	25.20	5.11	5.88
HuBERT Large	95.29	98.76	90.33	67.62	3.53	3.62	0.035	89.81	21.76	5.98	5.75
Gemma 4 12B	96.79	98.13	85.14	66.79	11.11	9.57	0.003	85.67	30.13	6.68	7.85
Inkling-Small	94.90	98.29	85.52	69.61	15.44	19.12	0.006	71.94	50.25	6.96	12.14
Gemma Distilled 800M	96.69	98.26	92.44	66.24	9.81	10.20	0.019	87.38	28.09	5.87	7.12
Gemma Distilled 400M	96.62	97.63	90.55	65.55	9.85	11.08	0.016	85.65	30.41	6.19	7.54
Gemma Distilled 200M	97.18	95.25	86.26	66.93	10.38	12.02	0.018	86.63	28.68	6.20	7.89
Gemma Distilled 100M	96.92	96.73	81.67	65.40	11.31	13.03	0.016	85.51	31.16	6.77	9.92

KS: Keyword Spotting, IC: Intent Classification, SID: Speaker Identification, ER: Emotion Recognition, PR: Phoneme Recognition, ASR: Automatic Speech Recognition w/o LM, SF: Slot Filling, ASV: Automatic Speaker Verification, SD: Speaker Diarization.

are jointly encoded, and the green dashed line indicates the 40.4% baseline accuracy achieved by Gemma-4 12B under the same condition. We observe that the performance of both models stays relatively constant around their baseline performance until Layer 12, at which point it begins to drop—degrading to chance level (25%) in the case of Gemma-4 12B, and dropping to approximately 37% for Inkling-Small. On the surface, these results are consistent with our observations of Gemma-4 12B in Section 3.1. We saw significantly higher context sensitivity in the model's later layers compared to Inkling-Small, and here we see a correspondingly large drop in OmniBench accuracy when merging activations at these later layers.

Conversely, our analysis in the previous section suggested that cross-modal context has very small impact on Inkling-Small's audio embeddings (Fig. 1) and image embeddings (Fig. 2(b)). This seems contrary to its performance on OmniBench, where we see a marked degradation in the independently-embedded modalities after Layer 12 (although not as severe as Gemma-4). To explain this discrepancy, we computed the linear CKA similarity scores on Inkling's segment-wise embeddings between the jointly and independently embedded scenarios, as shown in Fig. 4. Here, a segment refers to either *content* tokens (audio, image, question) or *boundary* tokens. While the similarity scores for the content segments remain high throughout the layers, the boundary tokens see a sharp decline after Layer 12. This suggests that Inkling-Small stores its cross-modal context primarily in the boundary tokens between the modality token segments.

3.3 Performance As An Audio Encoder

Although our previous experiments have provided evidence that the early layers of both Gemma-4 12B and Inkling-Small do behave as image and audio encoders, the question arises as to whether

Table 2: Frozen layer-10 image representations of Gemma-4 12B compared to dedicated vision encoders on the MIEB-Lite image-only tasks. Classification reports accuracy (Acc) and clustering reports Normalized Mutual Information (NMI). Best per task in **bold**.

	Image Classification (Acc ↑)								**Clustering** (NMI ↑)		
Model	Country211	DTD	EuroSAT	GTSRB	OxfordPets	PatchCamelyon	RESISC45	SUN397	ImageNetDog15	TinyImageNet	**Avg**
SigLIP2-base	**18.05**	**79.48**	**86.40**	**80.76**	92.92	**73.52**	**88.36**	**77.16**	79.61	59.82	**73.61**
DINOv2-base	10.73	75.93	81.66	53.48	**94.12**	72.73	81.30	71.03	**85.67**	**74.31**	70.10
Gemma-4 12B (L10)	10.10	71.90	83.20	48.10	53.70	72.70	78.10	56.40	50.40	59.30	58.39

these learned representations can generalize across modality-specific tasks. To that end, we evaluate how well the frame-level audio embeddings computed by both models perform in various downstream speech processing tasks using SUPERB (Yang et al., 2021), which is a widely-used evaluation benchmark for assessing the performance of pre-trained speech embedding models across a range of tasks including speech recognition, emotion recognition, speaker identification, and more. We compare our performance with the numbers reported for both the Base and Large variants of the HuBERT (Hsu et al., 2021) and wav2vec 2.0 (Baevski et al., 2020) SSL models on SUPERB. We also include the reported log mel filterbank features (FBANK) as an additional baseline since they are a popular feature used in many speech tasks.

Table 1 separates the tasks into two categories: those that use a single mean-pooled embedding per layer and those that use all of the frame-wise embeddings per-layer. The embeddings of the first 12-layers are then combined using a weighted-sum to get the final output. The tasks that use mean-pooled embeddings are all classification tasks and measure accuracy. We see that for both encoder-free models, the performance on these tasks is competitive with the established baselines. This demonstrates that the early layers of both models function as audio encoders for purposes of these tasks, producing global features that generalize across different speech classification tasks.

On the tasks that require more granular representations, such as phoneme and speech recognition, we surprisingly see that the frame-wise embeddings under-perform relative to the baseline SSL models on tasks such as ASR and phone recognition. This is consistent with the ASR results reported by Ding (2026), which measured Gemma-4 12B's ASR performance to be substantially worse than the Gemma-4 2B and 4B models - for example, the 12B model reaches a WER of 23.9% on GigaSpeech, while the 2B and 4B models respectively reach 11.8% and 11.1% on the same task.

3.4 Performance as a vision encoder

Similarly, we evaluate whether the representation learned in these early layers for both Gemma-4 12B and Inkling-Small behave as a vision encoder. We use the image-only tasks found in the lite subset of the MIEB Xiao et al. (2025) benchmark. This gives us 10 tasks total, split into 8 classification and 2 clustering tasks. For Gemma-4 12B, we probe over the first 12 layers, as shown in Figure 6 and Figure 7 (See Appendix Sec. A.1), and find that the representation provided by the 10th layer gives the highest overall score on the given tasks. Table 2 shows that performance on clustering underperforms compared to the two vision encoders. In classification tasks, it is competitive on half the tasks, while it is underperforming on the others.

3.5 Layer-wise analysis of audio representations

Our next analysis treats the early layers of Gemma-4 12B and Inkling-Small as audio encoders, and probes what kind of information is represented at each layer. We closely follow the work of Pasad et al. (2021; 2023), who performed layer-wise analysis of the information (acoustic, phonetic, lexical, semantic, etc.) present in the different layers of self-supervised speech encoders such as HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and wav2vec2.0 (Baevski et al., 2020). In Pasad et al. (2021; 2023), the authors show that while these models do exhibit meaningful differences from one another, a common pattern in their representations does emerge: the lower levels of

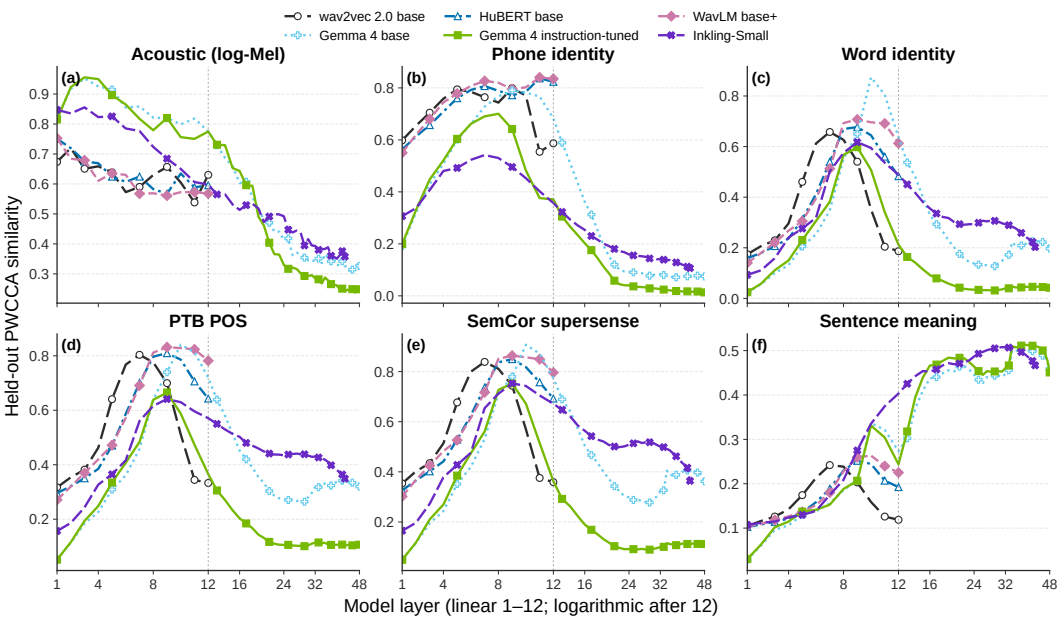


Figure 5: Layer-wise analysis of the audio token representations learned by Gemma-4 12B, Inkling-Small, and several self-supervised baselines.

these models tend to capture low-level acoustic information, the middle to later layers capture phonetic information, and the layers after that capture information about word identity and meaning.

We replicated the analysis pipeline of (Pasad et al., 2021; 2023) and applied it to both Gemma-4 12B and Inkling-Small. Specifically, we used projection-weighted canonical correlation analysis (PWCCA) (Morcos et al., 2018) to quantify the similarity between the layer-wise representations of each model and acoustic features (via log-mel featurebank features), phones (via one-hot phonetic labels aligned at the frame level), words (via one-hot word labels aligned at the frame level), syntax (via Penn Treebank (Marcus et al., 1993) part-of-speech tag labels aligned with words), semantics (via SemCor (Miller et al., 1993) word-level supersenses). We also add an additional comparison to whole-sentence meaning via sentence embeddings computed using the MPNet text embedding model (Song et al., 2020) applied to the text transcripts of each utterance.

Figure 5 shows the results of this analysis on Gemma-4 12B, Inkling-Small, HuBERT-Base, WavLM-Base+, and wav2vec2.0-Base. Regarding the similarity to log-mel filterbank features, we observe that Gemma-4 12B starts with a very low score at layer 0 (representing the raw waveform samples) but rapidly peaks at layer 3 with a PWCCA similarity score of 0.95. This indicates that although Gemma-4 12B takes as input windows of raw waveform sample amplitudes, the first 3 layers of the model learn a transformation of those features that nearly perfectly captures log-mel filterbank energies (although this does not rule out the possibility that the layers encode additional information beyond that). It should be noted that the layer 0 scores for the HuBERT, WavLM, and wav2vec2.0 models with the log-mel filterbank energies already starts at a high value, which we attribute to the fact that these models all use several convolutional layers as a waveform encoder, and the layer 0 features represent the output of this encoder. Inkling-Small takes as input dMel features, so it is not surprising that its initial layers have a high similarity to log-Mel filterbank energies.

Regarding the similarity of Gemma-4 12B and Inkling-Small features to phonetic and word identity, we observe that phonetic similarity peaks at layer 8 for Gemma and layer 7 for Inkling, and lexical similarity peaks in both models at layer 9. Both of these patterns are similar to what is observed in the self-supervised baselines, except for the fact that the phonetic information in the self-supervised models stays relatively high after its initial peak. Conversely, both phonetic and word information in Gemma-4 12B and Inkling-Small drop rapidly after their peaks, raising the question of what information is being captured by the models at layer 10 and onward.

In the bottom of Figure 5, we try to answer this question by measuring the similarity of Gemma-4 12B's and Inkling-Small's representations to higher levels of abstraction, starting with part of speech tags and word meanings (left and center plots). However, we find that both of these similarities peak alongside word identity, before rapidly falling. However, when we examine the similarity to whole-sentence embeddings we see this information begin to steadily grow at layer 10, and remain high throughout the remainder of the model. Overall, our analysis suggests that both Gemma-4 12B and Inkling-Small evolve their speech audio representations by first capturing low-level acoustic information, mapping this to phonetic sequences followed by word sequences, and then finally mapping these word sequences to representations that capture higher-level sentence meaning.

3.6 Isolating Implicit Encoders via Distillation

Throughout the previous several sections we have shown that the early layers of Gemma-4 12B and Inkling-Small behave like encoders that process each modality independently. However, in both of these models the same set of parameters serves every modality. We ask whether the encoding functions of each modality are separable into dedicated models, and how much of the model's capacity a single modality actually requires. We focused on the audio modality, and used distillation to train a dedicated compact audio encoder from the first 12 layers of the pretrained Gemma-4 12B model. We trained randomly initialized student models of various sizes, ranging from 800M down to 100M parameters. Every student model maintained a similar architecture to that of the teacher model, preserving the 12 blocks, the query head and key/value head count, and the same attention pattern, including the alternation sliding-window at the sixth and twelfth layers. We resized the student models by scaling down the model dimension, as well as the dimension of the attention heads.

Our distillation process used approximately 8000 hours of audio sampled from GigaSpeech (Chen et al., 2021), Common Voice English (Ardila et al., 2020), and AudioSet (Gemmeke et al., 2017). We performed layer-to-layer distillation using the DistilHuBERT objective (Chang et al., 2022). We evaluated the quality of the distilled audio encoders on the SUPERB benchmark, the results of which can be found in Table 1. Distillation largely preserves the teacher's performance even at a student model roughly 7% of the teacher's size, suggesting that the intrinsic capacity of the implicit audio encoder in Gemma-4 12B is much less than the 2.73B parameters of the first 12 layers.

4 Discussion

In this paper, we analyzed the behavior of Gemma-4 12B and Inkling-Small, two recently released encoder-free multimodal LLMs. We hypothesized that the early layers of these models would learn to function similarly to traditional modality-specific encoders found in other multimodal LLMs, and conducted a series of experiments whose results overall supported this hypothesis. We tested these implicit encoders the SUPERB and MIEB benchmarks, and found that on some tasks they are competitive with existing encoders, but achieve underwhelming performance on other tasks. We also analyzed the layer-wise representations both LLMs when used as a speech audio encoder, and found several similarities to existing speech encoders. Finally, we attempted to isolate the implicit speech audio encoder formed by the first 12 layers of Gemma-4 12B via distillation and found that we could retain most of its performance using as little as 7% of its parameters.

Encoder-free multimodal LLMs are appealing to researchers and developers because they offer a simplified, monolithic model architecture and training recipe that does not need to be assembled from multiple, separately-trained components. However, this could potentially come with costs. As our analysis has shown, these models are very likely to "re-invent the wheel" each time they are trained from scratch, by spending model parameters and training time re-learning much of the same functionality that a pre-trained encoder could offer off-the-shelf. It is not yet clear whether from a task performance perspective encoder-free models have other inherent advantages over multimodal LLMs built with dedicated encoders. Future work along this direction should try to further quantify what advantages and disadvantages encoder-free architectures have over using separate encoders and LLM backbones in terms of task performance and training requirements (both data amount and compute time). We conjecture that it may be possible to achieve the best of both worlds, for example by initializing the early layers of an encoder-free architecture with parameters from pre-trained encoder models, or by using knowledge distillation from pre-trained encoders into the early layers of an encoder-free LLM as an additional training objective to speed up learning.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in experimental implementation, drafting figures, literature review and searching for related work, and editing/polishing writing. We have not used generative AI tools for any other tasks with required disclosure. We have reviewed all AI-assisted work, by manually checking suggested references and for the correctness of AI-generated experimental code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper analyzes the representations learned by open-weight models. We do not believe that it introduces any novel ethical hazards beyond those already existing with the models being studied.

REPRODUCIBILITY STATEMENT

We do not introduce new models, and only analyze existing open-weight models. While we believe that our experimental analysis could be easily reproduced with the information present in the paper, we intend to release the source code for our analysis.

REFERENCES

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Model card, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 4218–4222, 2020. URL https://arxiv.org/abs/1912.06670.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020. URL https://arxiv.org/abs/2006.11477.

Richard He Bai, Tatiana Likhomanenko, Ruixiang Zhang, Zijin Gu, Zakaria Aldeneh, and Navdeep Jaitly. dMel: Speech tokenization made simple, 2024. URL https://arxiv.org/abs/2407.15835.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing Fuyu-8B. https://www.adept.ai/blog/fuyu-8b, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7087–7091, 2022. URL https://arxiv.org/abs/2110.01900.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech*, pp. 3670–3674, 2021. URL https://arxiv.org/abs/2106.06909.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. URL https://arxiv.org/abs/2110.13900.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2407.06438.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models, 2024. URL https://arxiv.org/abs/2409.17146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://arxiv.org/abs/2406.11832.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. EVEv2: Improved baselines for encoder-free vision-language models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2502.06788.

James Ding. How good is Gemma 4 at ASR tasks? benchmarking against the open ASR leaderboard. Blog post, https://twango.dev/writing/gemma4-asr-benchmark, June 2026. URL https://twango.dev/writing/gemma4-asr-benchmark. Accessed: 2026-09-22.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Gemini Team, Google. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Google DeepMind. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. URL https://arxiv.org/abs/2106.07447.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Vedang Lad, Wes Gurnee, and Max Tegmark. Remarkable robustness of LLMs: Stages of inference? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2406.19384.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models, 2024. URL https://arxiv.org/abs/2409.15272.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2304.08485.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology (HLT)*, pp. 303–308. Association for Computational Linguistics, 1993.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://arxiv.org/abs/1806.05759.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2024. URL https://arxiv.org/abs/2410.07149.

OpenAI. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021. URL https://arxiv.org/abs/2107.04734.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2211.03929.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics (TACL)*, 12:372–391, 2024. URL https://arxiv.org/abs/2307.00162.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023. URL https://arxiv.org/abs/2212.04356.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics (TACL)*, 8: 842–866, 2020. URL https://aclanthology.org/2020.tacl-1.54/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016. URL https://aclanthology.org/P16-1162/.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1701.06538.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16857–16867, 2020. URL https://arxiv.org/abs/2004.09297.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.13289.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601, 2019. URL https://aclanthology.org/P19-1452/.

Thinking Machines Lab. Inkling: Our open-weights model. https://thinkingmachines.ai/news/introducing-inkling/, 2026. URL https://thinkingmachines.ai/news/introducing-inkling/. Model card: https://thinkingmachines.ai/model-card/inkling/.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15366–15394, 2024. URL https://aclanthology.org/2024.acl-long.820/.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22187–22198, 2025. doi: 10.1109/ICCV51701.2025.02060.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Proc. Interspeech*, pp. 1194–1198, 2021. URL https://arxiv.org/abs/2105.01051.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833. Springer, 2014. URL https://arxiv.org/abs/1311.2901.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. URL https://arxiv.org/abs/2303.15343.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google USM: Scaling automatic speech recognition beyond 100 languages, 2023. URL https://arxiv.org/abs/2303.01037.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2024. URL https://arxiv.org/abs/2411.18620.

A Appendix

A.1 Layer-Wise Results on MIEB

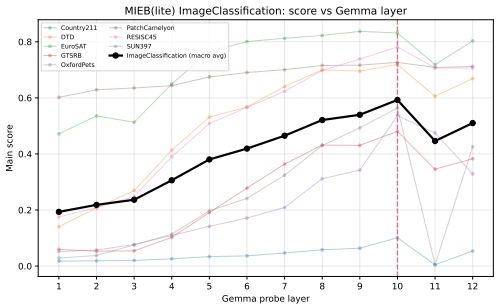


Figure 6: Per-layer sweep for Gemma4-12B on Image Classification Tasks

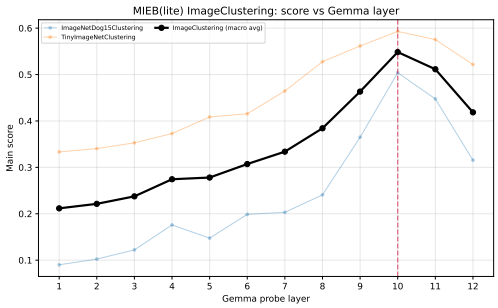


Figure 7: Per-layer sweep for Gemma4-12B on Image Clustering Tasks