Do Language Models Associate Sound with Meaning? A Multimodal Study of Sound Symbolism

Jinhong Jeong[*1], **Sunghyun Lee**[*1], **Jaeyoung Lee**[2], **Seonah Han**[3], **Youngjae Yu**[†2]

[1]Yonsei University
[2]Seoul National University
[3]Korea University
{jjhsnail0822, sheepswool}@yonsei.ac.kr, {jerry96, youngjaeyu}@snu.ac.kr, sunahan@korea.ac.kr

Abstract

Sound symbolism is a linguistic concept that refers to non-arbitrary associations between phonetic forms and their meanings. We suggest that this can be a compelling probe into how Multimodal Large Language Models (MLLMs) interpret auditory information in human languages. We investigate MLLMs' performance on phonetic iconicity across textual (orthographic and IPA) and auditory forms of inputs with up to 25 semantic dimensions (e.g., *sharp vs. round*), observing models' layer-wise information processing by measuring phoneme-level attention fraction scores. To this end, we present **LEX-ICON**, an extensive mimetic word dataset consisting of 8,052 words from four natural languages (English, French, Japanese, and Korean) and 2,930 systematically constructed pseudo-words, annotated with semantic features applied across both text and audio modalities. Our key findings demonstrate (1) MLLMs' phonetic intuitions that align with existing linguistic research across multiple semantic dimensions and (2) phonosemantic attention patterns that highlight models' focus on iconic phonemes. These results bridge domains of artificial intelligence and cognitive linguistics, providing the first large-scale, quantitative analyses of phonetic iconicity in terms of MLLMs' interpretability.

Code — https://github.com/jjhsnail0822/sound-symbolism
Extended version — https://arxiv.org/pdf/2511.10045

1 Introduction

Sound symbolism, which suggests that phonetic sounds and their meanings have a significant correlation, presents a cognitively grounded exception to the principle of linguistic arbitrariness (Hinton et al. 2006; Dingemanse et al. 2015). For instance, when people are presented with two images of sharp and round shapes called "kiki" and "bouba", the overwhelming majority of participants consistently match "kiki" with pointed shapes and "bouba" with round shapes, regardless of their cultural background (Köhler 1967; Ramachandran and Hubbard 2001). This iconicity effect illustrates the universal sensory associations of humans that intuitively link certain phonetic sounds with certain semantic features, which facilitate early-childhood or second-language acquisition (Imai and Kita 2014), and commercially nuanced brand naming (Yorkston and Menon 2004).

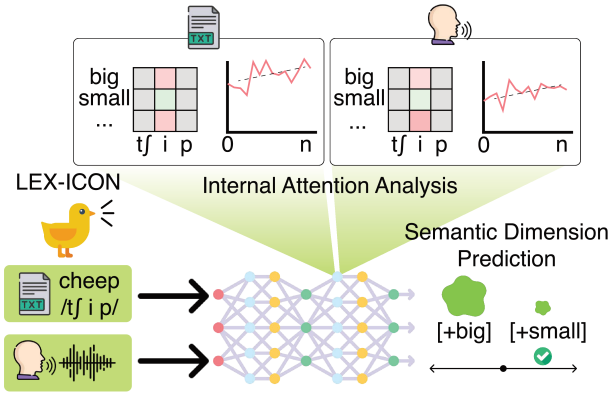

Figure 1: Phonetic iconicity investigation for MLLMs using natural and constructed mimetic words from text and audio modalities in LEX-ICON. We conduct quantitative evaluations for up to 25 semantic dimensions and examine layer-wise attention fraction scores to identify how phonemes and meanings are related within the models.

Recent advancements in Multimodal Large Language Models (MLLMs) that incorporate audio-modality input into an integrated representation space (OpenAI et al. 2024; Xu et al. 2025) shed light on new possibilities to systematically analyze human-like phonetic iconicity of language models. With MLLMs as a novel test bed, we formulate the following two key research questions:

[*]These authors contributed equally.
[†]Corresponding author. J.J. introduced sound symbolism and performed the semantic dimension prediction. S.L. conducted the internal attention analysis. J.L. introduced the idea of formulating the problem as multimodal interpretability. S.H. built the constructed word data and interpreted the linguistic implications.

RQ 1. How do MLLMs associate sound-symbolic words with semantic features similar to human phonetic intuition?

RQ 2. How do MLLMs' internal attention patterns align with phonosemantic relationships?

Mimetic words provide a compelling probe to address these subjects. Represented by onomatopoeias and ideophones[1], mimetic words refer to words in which the non-arbitrary association of sound form and meaning is prominent (Akita and Pardeshi 2019), such as "boom" or "whizz." We construct **LEX-ICON**, a dataset of natural mimetic words and constructed pseudo-words designed to maximize sound-symbolic effect. With LEX-ICON, we apply the semantic dimension paradigm, where word meanings are projected onto binary scales (e.g., *fast vs. slow*), to examine MLLMs' abilities to infer the meaning of words from their form (original text, IPA, audio) rather than their content. We further analyze models' internal representations to identify whether models actually focus on iconic phonemes, as illustrated in Figure 1.

Our experimental results show that MLLMs demonstrate phonetic intuitions across multiple semantic dimensions, for not only natural words but also constructed pseudo-words that exclude models' memorization (e.g., *sharp vs. round*). In particular, MLLMs exhibit modality-specific preferences across semantic dimensions, using audio for acoustically grounded features (e.g., *big vs. small*) and text for articulatory or visually driven features (e.g., *beautiful vs. ugly*). However, we reveal that there still remain discrepancies between the semantic dimensions where humans exhibit high scores and those where models perform well. Through phoneme-level attention analysis, we suggest that MLLMs attend to sound-symbolic phonemes corresponding to their meanings. In particular, we demonstrate that attention fraction to these sound-symbolic phonemes is more prominent in models' late-layers when processing constructed pseudo-words, suggesting that the relationship between phonemes and meanings can be systematically represented in the deep layers. Based on these findings, we summarize our key contributions as follows.

1. We conduct the first large-scale investigation of phonetic iconicity in MLLMs, grounded in LEX-ICON, a novel multilingual mimetic word dataset spanning multiple language families.
2. We provide a quantitative evaluation methodology of sound symbolism through the semantic dimension approach that contributes to linguistics by capturing phonetic iconicity akin to humans.
3. We present a comprehensive analysis that elucidates models' internal mechanisms towards sound symbolism by measuring phoneme-level attention scores, thereby contributing to the field of model interpretability.

Our approach integrates two distinct but interconnected domains, addressing both artificial intelligence and linguistics. For the former, we observe an integration of form and meaning within MLLMs using sound symbolism probes, revealing new insights in terms of model interpretability. From a linguistic perspective, we propose a substantiation for nonhuman intelligence's phonetic intuitions on mimetic word data, providing a quantitative basis for linguistic experiments that have been conducted mostly with humans.

[1]In this paper, "onomatopoeia" and "ideophone" refer to words that depict sounds and non-auditory sensory imagery, respectively.

2 Related Work

2.1 Sound Symbolism in Linguistics

Early modern linguists question the arbitrary relationship between signifier and signified, suggesting the phenomenon of intuitive phonetic symbolism in humans (Usnadze 1924; Sapir 1929; Köhler 1967). Sapir (1929)'s experiment has exhibited that people who are presented with unfamiliar object names "mil" and "mal" and are asked to estimate the size of the objects respond that the latter is larger than the former (Parise and Spence 2012). Recent studies continue to observe these phenomena (Hinton et al. 2006; Lockwood and Dingemanse 2015; Ćwiek et al. 2022), extending them by scaling to a large amount of data (Thompson, Van Hoey, and Do 2021; Winter et al. 2024) or measuring the exact semantic dimensions associated with each phoneme (Monaghan and Fletcher 2019; Sidhu, Vigliocco, and Pexman 2022; Sidhu 2025). For the mimetic words, experimental research has also demonstrated that people can infer a certain degree of meaning from the words in languages besides their mother tongue (Shinohara and Kawahara 2010; Dingemanse et al. 2016).

2.2 Phonetic Iconicity for LLMs

Iconicity tasks for language models have been used as a means of assessing whether the models have human-like phonetic intuition (Cai et al. 2024; Duan et al. 2024). Early studies focus on analyzing phonesthemic information contained in word embedding spaces (Abramova, Fernández, and Sangati 2013; Abramova and Fernández 2016). Recent studies demonstrate that the effect of non-arbitrary integration of linguistic form and meaning is substantiated in text-only LLMs (Miyakawa et al. 2024; Marklová et al. 2025), vision and image generation models (Loakman, Li, and Lin 2024; Alper and Averbuch-Elor 2024; Shinto and Iizuka 2024; Iida and Funakura 2024), and audio-visual models (Tseng et al. 2024), yet they are not extensive enough in scope to include exhaustive and multilingual mimetic word data or diverse semantic feature analysis.

2.3 Multimodal Interpretability

Recently, the field of model interpretability has witnessed significant growth (Elhage et al. 2021; Zou et al. 2025; Wang et al. 2023). Motivated by the rapid progress of MLLMs (OpenAI et al. 2024; Xu et al. 2025), there has been increasing interest in multimodal interpretability (Lin et al. 2025). For instance, Neo et al. (2025) ablates a subset of visual tokens to observe the resulting differences in model output, while Nikankin et al. (2025) demonstrates that models employ distinct processing circuits depending on the input modality. Despite these advances, prior work has predominantly focused on the visual modality, with less exploration

	En.	Fr.	Ja.	Ko.	Con.	Total
# Words	826	809	1418	4999	2930	**10982**

Table 1: Data distribution of LEX-ICON across *natural* (8,052 words from English, French, Japanese, and Korean) and *constructed* (2,930 words) groups. Japanese and Korean are known for their rich mimetic words (Hamano 1986; Kwon 2018), which contributes to their high proportion.

of audio-based interpretability. Yang et al. (2025b) investigates how models handle auditory input, but their analysis is limited to simple settings where the audio simply serves as a direct vocalization of textual data. In this work, we examine audio interpretability through the lens of sound symbolism, highlighting how phoneme-level features encode meaning in ways that are intrinsically tied to the auditory modality.

3 LEX-ICON

We build LEX-ICON, a dataset with *natural* word group consisting of existing large-scale mimetic words derived from four natural languages (English, French, Japanese, and Korean), as well as *constructed* word group comprising systematically generated pseudo-words. Table 1 and 2 summarize the distribution of the overall datasets. Figure 2 also illustrates an overall dataset construction flow.

3.1 Semantic Dimension

We employ the semantic dimension methodology, a concept originated from "semantic differential" by Osgood, Suci, and Tannenbaum (1957), which consists of two semantic feature adjectives located at opposite extremes, such as "big" and "small." By applying this methodology to sound symbolism experiments, we can simultaneously and precisely measure the multifactorial meanings contained in a single word. We annotate up to 25 pairs of predefined semantic features as per Sidhu, Vigliocco, and Pexman (2022) for each word in the LEX-ICON, which is the most diverse scale in terms of studies on LLMs' iconicity.

3.2 Natural Mimetic Words

Data Collection. We manually collect 8,052 mimetic words and definition data consisting of onomatopoeias and ideophones in English, French, Japanese, and Korean from specialized mimetic word lexicons and authoritative dictionaries for each language. We then extract the most representative definitions of these mimetic words from dictionaries such as the Oxford English Dictionary (Simpson 1989), Le Petit Robert (Alain Rey 2022), Nihon Kokugo Daijiten (Shogakukan 2006), and the Standard Korean Language Dictionary (National Institute of Korean Language 2025). For further source information, see the Appendix.

Input Type Variation. To observe the effect of trained token memorization and modality change, we create three types of word form that contain textual and auditory input: (1) original text, (2) IPA-converted text with phoneme-level spacing, (3) audio waveform converted using text-to-speech (TTS) software. We perform IPA conversion with the Epitran package (Mortensen, Dalmia, and Littell 2018), and obtain the TTS dataset using Google Text-to-Speech (Google 2025) for English, French, and Japanese; and MeloTTS (Zhao, Yu, and Qin 2023) for Korean.

Dimension	Natural	Constructed	Total
good-bad	2083	–	2083
beautiful-ugly	929	462	1391
pleasant-unpleasant	3380	–	3380
strong-weak	4299	208	4507
big-small	2073	1687	3760
rugged-delicate	2664	–	2664
active-passive	3884	–	3884
fast-slow	2051	1437	3488
sharp-round	2323	1623	3946
realistic-fantastical	6883	501	7384
structured-disorganized	3712	–	3712
ordinary-unique	1585	208	1793
interesting-uninteresting	208	–	208
simple-complex	4322	1602	5924
abrupt-continuous	4703	1005	5708
exciting-calming	2256	1402	3658
hard-soft	3676	1136	4812
happy-sad	719	1463	2182
harsh-mellow	3106	1005	4111
heavy-light	2918	1341	4259
inhibited-free	2673	206	2879
masculine-feminine	378	1522	1900
solid-nonsolid	2255	893	3148
tense-relaxed	2956	206	3162
dangerous-safe	448	541	989
Total	**66484**	**18448**	**84932**

Table 2: Semantic dimension distribution of pseudo ground truth data by word group. We adopt 25 semantic dimension criteria by Sidhu, Vigliocco, and Pexman (2022) to annotate LEX-ICON. Six dimensions from the constructed word group are excluded by removing close-to-neutral data points. As a result, 19 dimensions remain for experiments in §4.

Large-Scale Annotation Process. To effectively generate a large-scale ground truth data, all natural language words and their definitions in the LEX-ICON are given four LLMs: GPT-4.1 (OpenAI 2025), Qwen3-32B (Yang et al. 2025a)[2], Gemma-3-27B (Team et al. 2025b), and Gemini-2.5-flash (Comanici et al. 2025). The models are prompted to annotate each semantic dimension of a word with one of a feature pair or neutral labels (e.g., selecting one of the "exciting", "calming", or "neither" option) for each word. See the Appendix for the detailed prompt.

We finalize the results as pseudo ground truth by selecting unanimously agreed features across all models, deleting "neither" labels to remove the meaningless features for each word. As shown in Table 2, this process filters 67.0% of the total annotation points, yielding 66,484 high-quality semantic features for natural words. In §4, we verify these pseudo ground truth data through a human evaluation ex-

[2]Non-reasoning mode.

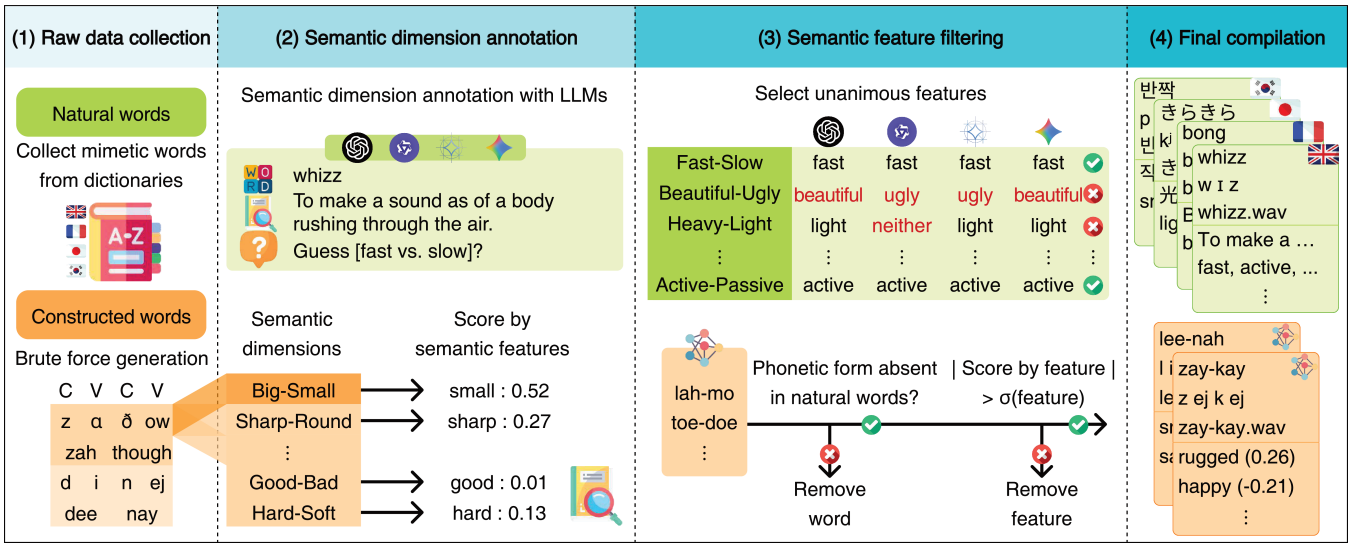

Figure 2: A comprehensive figure for the data construction flow of LEX-ICON. (1) We manually collect 8,052 mimetic words and definitions from dictionaries, and systematically construct 2,930 disyllabic pseudo-words. (2) Using four LLMs (GPT-4.1, Qwen3-32B, Gemma-3-27B, and Gemini-2.5-flash), we automatically annotate each word with semantic dimensions based on its definitions. (3) For natural words, we retain features agreed upon by all models. For constructed words, we filter out features that are close to neutral. (4) The final dataset contains 10,982 words with 84,932 semantic features with varied input types.

periment. For more information on the semantic dimension ground truth, refer to the Appendix.

3.3 Constructed Mimetic Words

Phoneme Combination Generation. We systematically construct novel words using a CVCV structure to create pseudo-words unlikely to be encountered during model training. We use 15 consonants from five categories: sonorants (/l/, /m/, /n/), voiced fricatives (/v/, /ð/, /z/), voiceless fricatives (/f/, /s/, /ʃ/), voiceless stops (/p/, /t/, /k/), and voiced stops (/b/, /d/, /g/). We use four vowels from two categories: front vowels (/i/, /ej/) and back vowels (/ɑ/, /ow/). After filtering against existing entries in the IPA-dict database (Doherty 2016) across four languages (US and UK English, French, Japanese, and Korean), we obtain 3,108 novel pseudo-words. We convert IPA symbols to English alphabet combinations for TTS compatibility and remove incorrectly pronounced words by Google TTS, yielding 2,930 final words.

Semantic Dimension Annotation. We utilize the empirical coefficients from Sidhu, Vigliocco, and Pexman (2022), who quantified associations between phoneme categories and semantic dimensions through human rating experiments. These coefficients represent how much phoneme categories deviate from the overall mean rating across 25 semantic dimensions. We assign corresponding scores to phonemes in our constructed words and calculate the mean score across the four phonemes. To focus on meaningful associations, we apply a threshold of 1.0 standard deviation from the neutral point, treating data points within this range and as semantically neutral, consistent with 'neither' labels used for multilingual mimetic words.

4 Semantic Dimension Prediction

We perform semantic dimension A/B tests to answer *RQ 1*, interpreting the results by semantic dimension, word group, and input type to quantitatively explore MLLMs' phonetic intuitions. Human evaluation is also conducted to ensure reliability of pseudo ground truth data in LEX-ICON, and the results are further compared to those of MLLM experiments. Details of the experiments are provided in the Appendix.

4.1 Experimental Settings

We employ MLLMs that officially support simultaneous text and audio inputs: Qwen2.5-Omni (Xu et al. 2025)[3], Gemini-2.5-flash (Team et al. 2025a), and GPT-4o (OpenAI et al. 2024)[4]. All experimental models cover the four languages and orthographies present in LEX-ICON. For the Qwen models, inference is performed on one RTX 4090 GPU. We apply a zero-shot prompting strategy with temperature set to 0 throughout all experiments to ensure reproducibility.

4.2 Methodology

For each word, we prompt MLLMs with binary questions on each of the semantic dimensions and measure macro-F1 scores for the results of each dimension. We provide the models with words in three input types, keeping the query part as text. We insert audio tokens at the given words' positions within the prompt so that the audio tokens have the same series of positional embeddings with the text tokens, ensuring that the models infer internal representations in an

[3] `Qwen2.5-Omni-3B` and `Qwen2.5-Omni-7B`.

[4] `gpt-4o` for text-only input, `gpt-4o-audio-preview` for audio-enabled input.

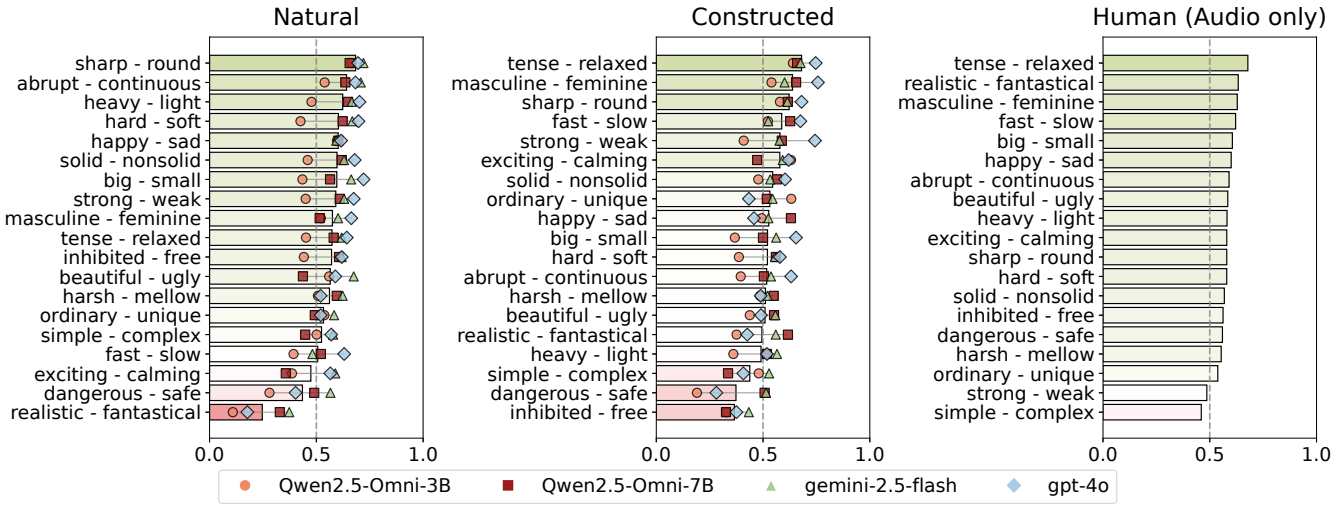

Figure 3: Macro-F1 score results for the semantic dimension A/B test. "Natural" and "Constructed" are results of LLM experiments, calculated by averaging all three input types (original text, IPA, and audio). Each dot represents each model's score for a given dimension. Human evaluation results only contain the "Audio" input type with sampled data for experimental feasibility, yet achieving superior scores compared to the baseline that demonstrate LEX-ICON's reliability.

Semantic Dimension Test
Given an IPA [WORD] with its pronunciation audio, which semantic feature best describes the word based on auditory impression? [WORD] ***b u m*** (AUDIO: <AUDIO>) [SEMANTIC DIMENSION] exciting vs. calming
[OPTIONS] **1: exciting** 2: calming Answer with the number only. (1-2)

Table 3: An example of prompts for the semantic dimension A/B test in §4.2. This example illustrates a case of a natural language group English word "boom" that both the IPA text tokens and audio tokens (<AUDIO>) are presented. Detailed prompts are provided in the Appendix.

integrated embedding space. An example prompt is illustrated in Table 3. The exact word input types are as follows:

- Original text tokens (e.g., "*boom*").
- Phoneme-level spaced IPA text tokens (e.g., "*b u m*").
- TTS audio tokens (e.g., <AUDIO>[5]).

We calculate macro-F1 scores to mitigate the imbalance in simple accuracy for each model across all combinations of word groups, input types, and semantic dimensions. Results for "natural" group are averaged equally across the four languages. Refer to the Appendix for detailed methods.

4.3 Result

Phonetic Intuition by Semantic Dimension. In Figure 3, MLLMs' macro-F1 scores averaged across all input types surpass the baseline score (0.50) in 84.2% (natural group) and 68.4% (constructed group) of the semantic dimensions, indicating that the models can detect phonetic iconicity not only in natural mimetic words that may have been memorized during training phase, but also in constructed pseudowords with maximized sound-symbolic effects. Overall performance becomes even larger when the comparatively small-scaled Qwen2.5-Omni-3B model is excluded. These results are supported by human evaluation results that score above the baseline in most dimensions, which guarantees the reliability of our pseudo ground truth data automatically annotated by LLMs from dictionary data in §3.1. Notably, the models' strong performance on the *sharp vs. round* dimension aligns with well-known cognitive linguistic experiments like the "bouba-kiki" effect (Ramachandran and Hubbard 2001).

Human-like Iconicity. Figure 4 shows that the Qwen2.5-Omni-7B model achieves the highest overall Pearson correlation coefficient with human evaluations across semantic dimensions, whereas larger models such as gemini-2.5-flash deviate more from human results. This suggests that while MLLMs can partially capture phonetic iconicity, they are still far from human-like semantic alignment. In particular, the relatively low correlation of IPA-converted natural word results indicates that linguistic arbitrariness from diverse languages may override iconic patterns even in mimetic words, as models' knowledge has been shaped by large-scale

[5] <AUDIO> represents a series of audio tokens that correspond to a given word.

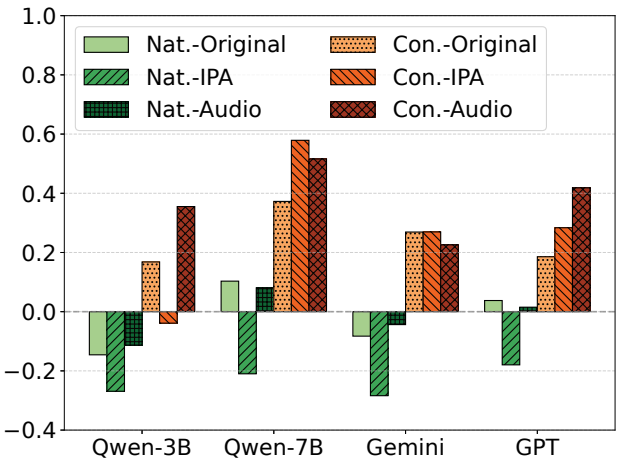

Figure 4: Pearson correlation scores with human evaluation results by word group and input type. Higher scores reflect greater similarity to humans' semantic dimension score distributions, where Qwen2.5-Omni-7B scores the highest correlation (maximum $r = 0.579$). In all models, constructed words elicit responses that are closer to human tendencies than natural words.

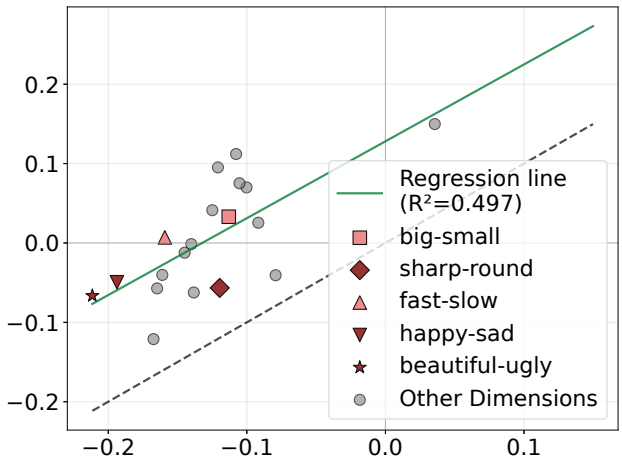

Figure 5: Advantage scores (macro-F1 differences) of audio inputs over the original text inputs by word group. X-axis indicates audio advantage scores for natural words, while Y-axis stands for constructed words. Each dot represents one semantic dimension, reflecting patterns aligned with linguistic implications, with an overall correlation (Pearson $r = 0.681$, Spearman $\rho = 0.705$).

distributional semantics of natural languages, possibly leading to a tendency to overlook subtle phonosemantic cues.

Linguistic Implication. Our findings reveal systematic modality preferences across semantic dimensions, providing computational evidence for the multi-mechanism nature of sound symbolism, as shown in Figure 5. For dimensions where acoustic features are theoretically central, such as size distinctions (*big vs. small*) that correlate with formant frequencies (Knoeferle et al. 2017) and speed distinctions (*fast vs. slow*) that relate to consonant voicing duration (Saji et al. 2013), the MLLMs show a pattern of enhanced performance when processing constructed words in audio format, consistent with the theoretical predictions. Conversely, for dimensions where non-acoustic mechanisms are proposed, such as shape associations (*sharp vs. round*) based on lip rounding gestures (Imai et al. 2025) and valence associations (*beautiful vs. ugly*, *happy vs. sad*) affected by articulatory properties (Körner and Rummer 2022), the models exhibit stronger reliance on textural representations.

5 Internal Attention Analysis

To address *RQ 2*, internal attention analysis focuses on the internal phenomena that emerge during the MLLM inference process. Then these phenomena are compared with linguistic theories to evaluate their correspondence.

5.1 Experimental settings

We utilize the Qwen2.5-Omni-7B model for the experiment due to its performance correlation most similar to humans, as well as its accessibility of model weights, which enables direct analysis of internal representations. All other settings for the experiment are as in §4 to maintain consistency.

5.2 Methodology

We calculate which semantic feature the model pays more attention to, for each IPA symbol given a semantic dimension. We perform layer-wise analyses of relative attention scores for each IPA symbol across contrasting semantic features, distinguishing by input types (IPA text and audio waveform) and word groups (natural and constructed words). Unlike original text tokens, the IPA text and audio input types facilitate phoneme-level investigation, as each phoneme is represented by at least one token.

Attention Fraction Score Extraction. For each inference conducted on the binary questions described in §4, we obtain the attention scores only when the model generates the correct response. For each layer, we calculate attention scores between tokens corresponding to a single IPA symbol (e.g., /w/ from /w ɪ z/) and tokens corresponding to each semantic feature (e.g., *fast vs. slow*). After retrieving scores, we normalize the paired scores so that the sum of the two semantic feature scores is equal to one. These fraction scores undergo head-wise and word-wise averaging within each layer to yield a single mean score per layer. To mitigate the bias of assigning higher attention scores to preceding tokens, we repeat each experiment with the semantic features presented in reversed order and compute the mean normalized attention fraction score across both feature-order conditions.

Token-IPA Symbol Alignment. When an IPA symbol spans multiple text tokens, the procedure first sums the attention scores across those tokens before fraction normalization. For audio inputs, we employ Montreal Forced Aligner (McAuliffe et al. 2025) to segment each audio waveform into phonemes over time. The phoneme sequence is then aligned with the model's 40 ms sampling period, with

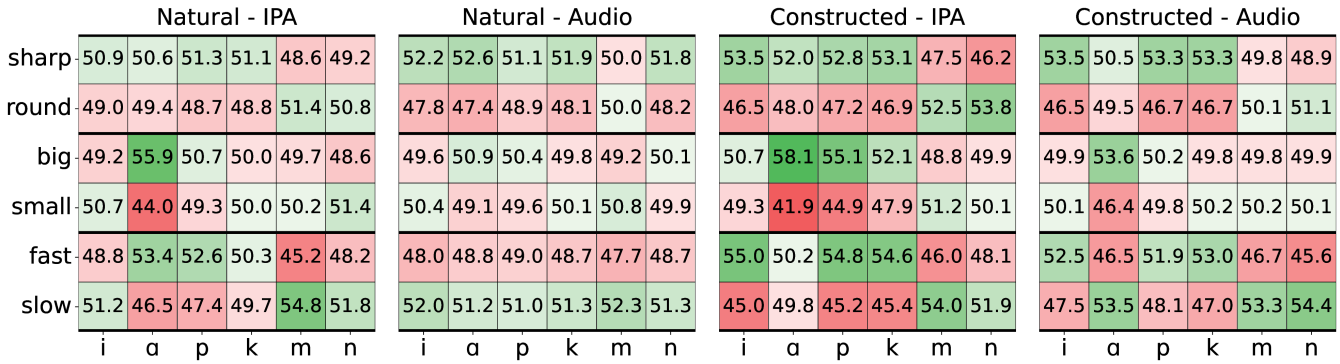

Figure 6: Attention fraction scores indicating the ratio of semantic dimensions that the model focuses on for each IPA, by word group and input type. For each IPA and semantic dimension, the model tends to attend more to semantic features that exhibit stronger associations with the given phonemes in linguistics. For instance, the model mostly associates *sharp* semantic feature with /p/ and /k/, *round* with /m/ and /n/, *big* with /ɑ/, and *small* with /i/ (Köhler 1967; Parise and Spence 2012).

consecutive occurrences of the same phoneme treated as a single IPA symbol. Further details are in the Appendix.

5.3 Result

An attention fraction score above 0.5 for a given IPA-semantic feature pair indicates the model's preferential focus on phonemes with sound-symbolic associations. For more details about the experiment, refer to the Appendix.

Layer-wise Attention Fraction Score. Figure 7 demonstrates that, for constructed words, the attention fraction scores for IPA text consistently exceed those for audio input type across layers (IPA = 0.523, audio = 0.506 on average), showing an upward trend toward the late layers. These lower phoneme-level attention scores on audio inputs may imply that multimodal models derive greater benefits from extensively trained texts than from the acoustic properties of less-trained audio data. On the other hand, the average attention fraction scores for natural words (IPA = 0.507, audio = 0.501 on average) are lower than those for constructed words. This phenomenon may occur because arbitrary form–meaning mappings of natural words attenuate phonosemantic cues, thereby obscuring iconic phonemes. This interpretation is further corroborated by Figure 4, which shows the low correlation between human evaluation scores for natural words in both IPA text and audio modalities.

Phoneme-Semantic Feature Relation. Figure 6 presents heatmaps of attention fraction scores for canonical IPA symbols and semantic dimensions by input type and word group. The patterns for constructed words in the IPA modality closely mirror prior findings in sound symbolism research. For example, phonemes such as /p/ and /k/ exhibit elevated attention under the *sharp* feature, whereas /m/ and /n/ associate with the *round* feature (Köhler 1967).

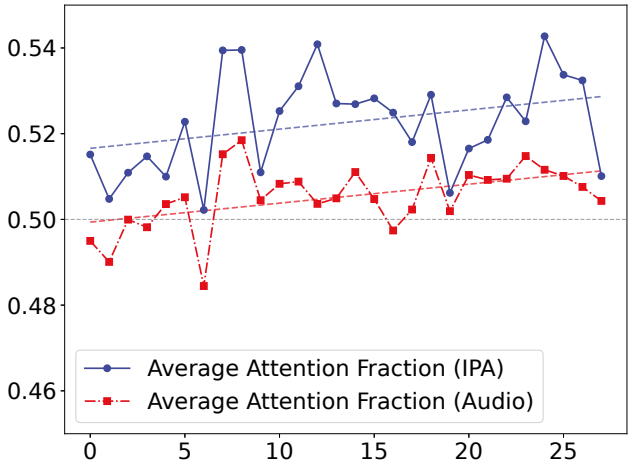

Figure 7: Attention fraction scores for the constructed word group, averaged across semantic dimensions. X-axis indicates the model's attention layer number, where Y-axis means the ratio at which the model attend to the correct semantic feature for a phoneme. While most layers score above the baseline fraction (0.50), the model focus on IPA text more than audio input for given iconic phonemes. Furthermore, the model tend to concentrate more on iconic phonemes in its late-layers.

6 Conclusion

In this work, we investigate MLLMs' phonetic iconicity on natural and constructed words via semantic dimension and internal attention analysis, constructing LEX-ICON, a large-scale mimetic word dataset for MLLM analysis for the first time. We discover that MLLMs have the ability to detect sound symbolism in both natural and constructed mimetic words, and pay a higher rate of attention in the internal layers to iconic phonemes. These results suggest that the models' sound-meaning association can be explained in terms of interpretability. Future work could further deepen the analytical methodology presented in this work through experiments with more human participants, investigate modality-specific information such as intonation, or extend it to application fields such as language learning or brand effects.

Acknowledgments

This work was partly supported by an Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University), No.RS-2025-02263598, Development of Self-Evolving Embodied AGI Platform Technology through Real-World Experience), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00354218, RS-2024-00353125). We express special thanks to KAIT GPU project. The ICT at Seoul National University provides research facilities for this study.

References

Abramova, E.; and Fernández, R. 2016. Questioning Arbitrariness in Language: a Data-Driven Study of Conventional Iconicity. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 343–352. San Diego, California: Association for Computational Linguistics.

Abramova, E.; Fernández, R.; and Sangati, F. 2013. Automatic labeling of phonesthemic senses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Akita, K.; and Pardeshi, P. 2019. *Ideophones, mimetics and expressives*, volume 16. John Benjamins Publishing Company.

Alain Rey, J. R.-D., ed. 2022. *Le Petit Robert de la Langue Francaise Dictionnaire 2023*. Paris: Le Robert, french edition edition.

Alper, M.; and Averbuch-Elor, H. 2024. Kiki or Bouba? Sound Symbolism in Vision-and-Language Models. arXiv:2310.16781.

Cai, Z.; Duan, X.; Haslett, D.; et al. 2024. Do large language models resemble humans in language use? In Kuribayashi, T.; Rambelli, G.; Takmaz, E.; Wicke, P.; and Oseki, Y., eds., *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 37–56. Bangkok, Thailand: Association for Computational Linguistics.

Comanici, G.; Bieber, E.; Schaekermann, M.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Ćwiek, A.; Fuchs, S.; Draxler, C.; et al. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841): 20200390.

Dingemanse, M.; Blasi, D. E.; Lupyan, G.; et al. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, 19(10): 603–615.

Dingemanse, M.; Schuerman, W.; Reinisch, E.; et al. 2016. What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, 92(2): e117–e133.

Doherty, L. 2016. IPA Dict: Monolingual Wordlists with Pronunciation Information in IPA. https://github.com/open-dict-data/ipa-dict. Accessed 2025-12-08.

Duan, X.; Xiao, B.; Tang, X.; and Cai, Z. G. 2024. HLB: Benchmarking LLMs' Humanlikeness in Language Use. arXiv:2409.15890.

Elhage, N.; Nanda, N.; Olsson, C.; et al. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Google. 2025. Google Cloud Text-to-Speech. https://cloud.google.com/text-to-speech. Accessed: 2025-06-25.

Hamano, S. S. 1986. *The sound-symbolic system of Japanese (ideophones, onomatopoeia, expressives, iconicity)*. University of Florida.

Hinton, L.; et al. 2006. *Sound symbolism*. Cambridge University Press.

Iida, H.; and Funakura, H. 2024. Investigating Iconicity in Vision-and-Language Models: A Case Study of the Bouba/Kiki Effect in Japanese Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Imai, M.; and Kita, S. 2014. The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical transactions of the Royal Society B: Biological sciences*, 369(1651): 20130298.

Imai, M.; Kita, S.; Akita, K.; et al. 2025. Does sound symbolism need sound?: The role of articulatory movement in detecting iconicity between sound and meaning. *The Journal of the Acoustical Society of America*, 157(1): 137–148.

Knoeferle, K.; Li, J.; Maggioni, E.; and Spence, C. 2017. What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, 7: 5562.

Köhler, W. 1967. Gestalt psychology. *Psychologische forschung*, 31(1): XVIII–XXX.

Körner, A.; and Rummer, R. 2022. Articulation contributes to valence sound symbolism. *Journal of Experimental Psychology: General*, 151(5): 1107.

Kwon, N. 2018. Iconicity correlated with vowel harmony in Korean ideophones. *Laboratory Phonology*, 9(1).

Lin, Z.; Basu, S.; Beigi, M.; et al. 2025. A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models. arXiv:2502.17516.

Loakman, T.; Li, Y.; and Lin, C. 2024. With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2849–2867. Miami, Florida, USA: Association for Computational Linguistics.

Lockwood, G.; and Dingemanse, M. 2015. Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, 6: 1246.

Marklová, A.; Milička, J.; Ryvkin, L.; et al. 2025. Iconicity in Large Language Models. arXiv:2501.05643.

McAuliffe, M.; et al. 2025. MontrealCorpusTools/Montreal-Forced-Aligner: Version 3.3.4.

Miyakawa, Y.; Matsuhira, C.; Kato, H.; et al. 2024. Do LLMs Agree with Humans on Emotional Associations to Nonsense Words? In Kuribayashi, T.; Rambelli, G.; Takmaz, E.; Wicke, P.; and Oseki, Y., eds., *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 81–85. Bangkok, Thailand: Association for Computational Linguistics.

Monaghan, P.; and Fletcher, M. 2019. Do sound symbolism effects for written words relate to individual phonemes or to phoneme features? *Language and Cognition*, 11(2): 235–255.

Mortensen, D. R.; Dalmia, S.; and Littell, P. 2018. Epitran: Precision G2P for Many Languages. In chair), N. C. C.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; and Tokunaga, T., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

National Institute of Korean Language. 2025. Standard Korean Language Dictionary [Online]. https://stdict.korean.go.kr/. Accessed: 2025-03-06.

Neo, C.; Ong, L.; Torr, P.; et al. 2025. Towards Interpreting Visual Information Processing in Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

Nikankin, Y.; Arad, D.; Gandelsman, Y.; et al. 2025. Same Task, Different Circuits: Disentangling Modality-Specific Mechanisms in VLMs. arXiv:2506.09047.

OpenAI; :; Hurst, A.; Lerer, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-06-25.

Osgood, C. E.; Suci, G. J.; and Tannenbaum, P. H. 1957. *The measurement of meaning*. 47. University of Illinois press.

Parise, C. V.; and Spence, C. 2012. Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, 220: 319–333.

Ramachandran, V. S.; and Hubbard, E. M. 2001. Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies*, 8(12): 3–34.

Saji, N.; Akita, K.; Imai, M.; et al. 2013. Cross-linguistically shared and language-specific sound symbolism for motion: An exploratory data mining approach. In *Proceedings of the annual meeting of the cognitive science society*, volume 35.

Sapir, E. 1929. A study in phonetic symbolism. *Journal of experimental psychology*, 12(3): 225.

Shinohara, K.; and Kawahara, S. 2010. A cross-linguistic study of sound symbolism: The images of size. In *Annual meeting of the berkeley linguistics society*, 396–410.

Shinto, R.; and Iizuka, H. 2024. Analyzing the Sensibility of Visual Language Models Using an Evolving Image Generation System: Focusing on Color Impressions and Sound Symbolism. In *Artificial Life Conference Proceedings 36*, volume 2024, 8. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Shogakukan. 2006. *Seisenban Nihon kokugo daijiten*. Seisenban Nihon kokugo daijiten. Shogakukan. ISBN 9784095210216.

Sidhu, D. M. 2025. Sound Symbolism in the Lexicon: A Review of Iconic-Systematicity. *Language and Linguistics Compass*, 19(1): e70006.

Sidhu, D. M.; Vigliocco, G.; and Pexman, P. M. 2022. Higher order factors of sound symbolism. *Journal of Memory and Language*, 125: 104323.

Simpson, E., Ja & Weiner. 1989. Oxford english dictionary. 3.

Team, G.; Anil, R.; Borgeaud, S.; et al. 2025a. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Team, G.; Kamath, A.; Ferret, J.; et al. 2025b. Gemma 3 Technical Report. arXiv:2503.19786.

Thompson, A. L.; Van Hoey, T.; and Do, Y. 2021. Articulatory features of phonemes pattern to iconic meanings: evidence from cross-linguistic ideophones. *Cognitive Linguistics*, 32(4): 563–608.

Tseng, W.-C.; Shih, Y.-J.; Harwath, D.; et al. 2024. Measuring Sound Symbolism In Audio-Visual Models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 1165–1172. IEEE.

Usnadze, D. 1924. Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung. *Psychologische Forschung*, 5: 24–43.

Wang, K. R.; Variengien, A.; Conmy, A.; et al. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *The Eleventh International Conference on Learning Representations*.

Winter, B.; Lupyan, G.; Perry, L. K.; et al. 2024. Iconicity ratings for 14,000+ English words. *Behavior research methods*, 56(3): 1640–1655.

Xu, J.; Guo, Z.; He, J.; et al. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Yang, A.; Li, A.; Yang, B.; et al. 2025a. Qwen3 Technical Report. arXiv:2505.09388.

Yang, C.-K.; Ho, N.; Lee, Y.-J.; and yi Lee, H. 2025b. AudioLens: A Closer Look at Auditory Attribute Perception of Large Audio-Language Models. arXiv:2506.05140.

Yorkston, E.; and Menon, G. 2004. A sound idea: Phonetic effects of brand names on consumer judgments. *Journal of consumer research*, 31(1): 43–51.

Zhao, W.; Yu, X.; and Qin, Z. 2023. MeloTTS: High-quality Multi-lingual Multi-accent Text-to-Speech.

Zou, A.; Phan, L.; Chen, S.; et al. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.