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

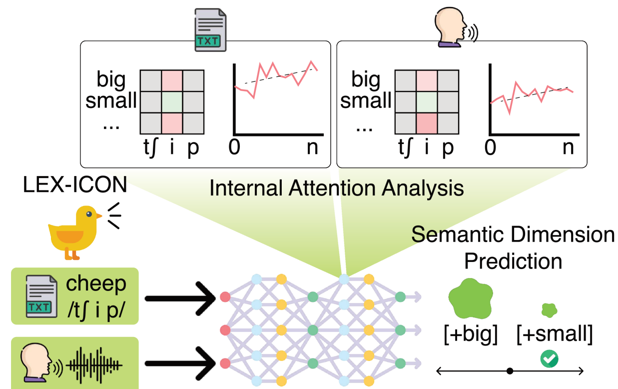

Figure 1: Phonetic iconicity investigation for MLLMs using natural and constructed mimetic words from text and audio modalities in LEX-ICON. We conduct quantitative evaluations for up to 25 semantic dimensions and examine layer-wise attention fraction scores to identify how phonemes and meanings are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

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).

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.

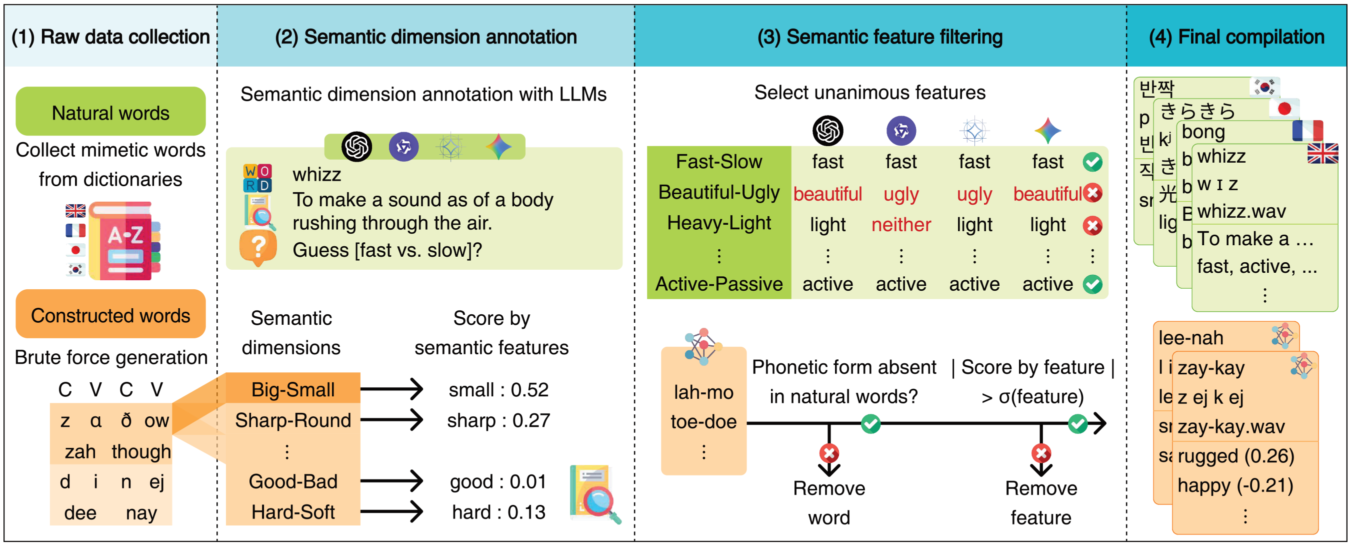

Figure 2: A comprehensive figure for the data construction flow of LEX-ICON. (1) We manually collect 8,052 mimetic words and definitions from dictionaries, and systematically construct 2,930 disyllabic pseudo-words. (2) Using four LLMs (GPT-4.1, Qwen3-32B, Gemma-3-27B, and Gemini-2.5-flash), we automatically 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.

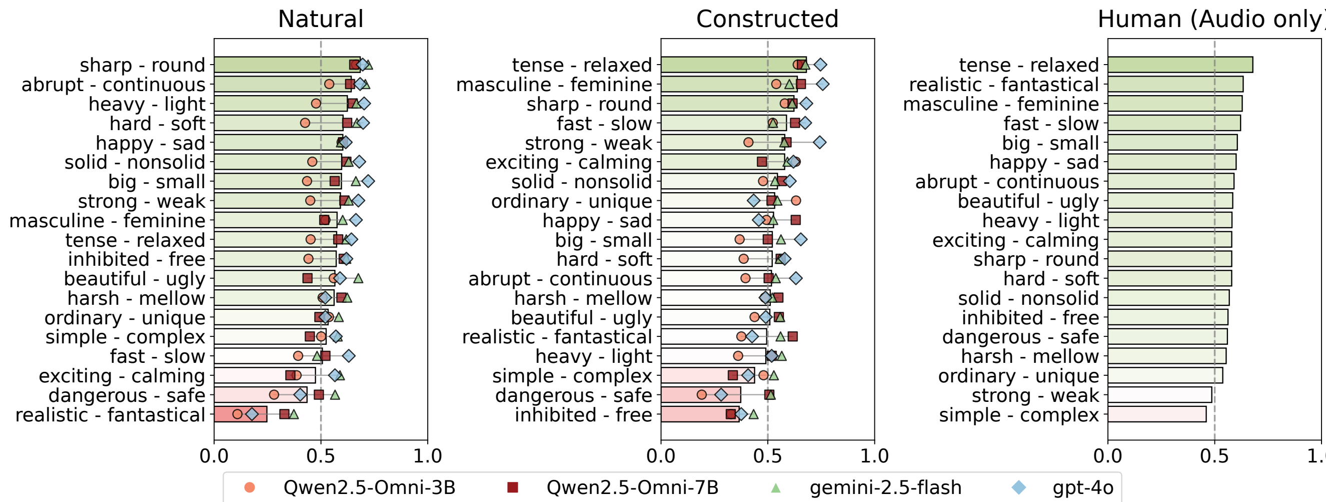

Figure 3: Macro-F1 score results for the semantic dimension A/B test. "Natural" and "Constructed" are results of LLM experiments, calculated by averaging all three input types (original text, IPA, and audio). Each dot represents each model's score for a given dimension. Human evaluation results only 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 pseudo-words 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.

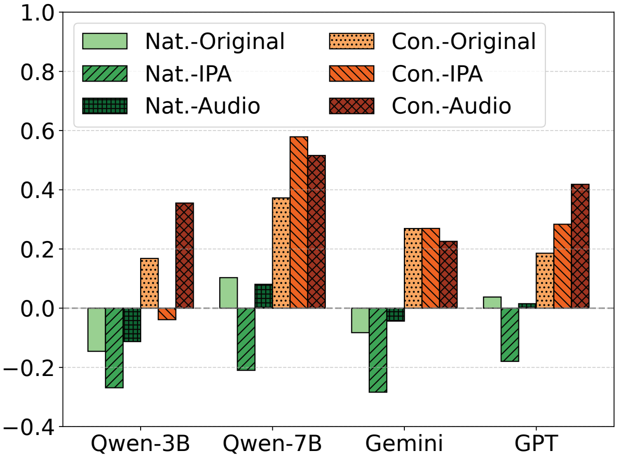

Figure 4: Pearson correlation scores with human evaluation results by word group and input type. Higher scores reflect greater similarity to humans' semantic dimension score distributions, where Qwen2.5-Omni-7B scores the highest correlation (maximum $r = 0.579$). In all models, constructed words elicit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

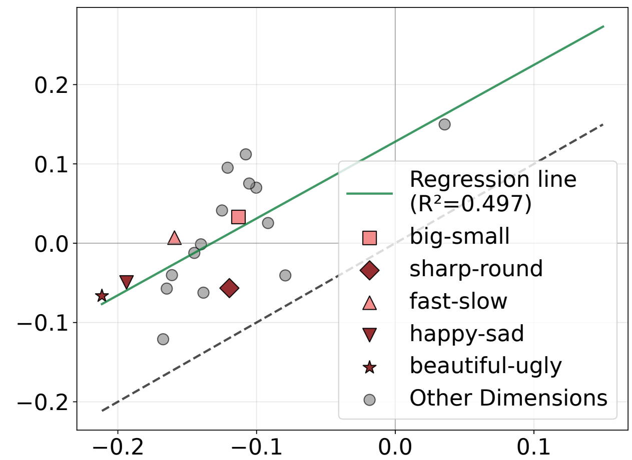

Figure 5: Advantage scores (macro-F1 differences) of audio inputs over the original text inputs by word group. X-axis indicates audio advantage scores for natural words, while Y-axis stands for constructed words. Each dot represents one semantic dimension, reflecting patterns aligned with linguistic 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

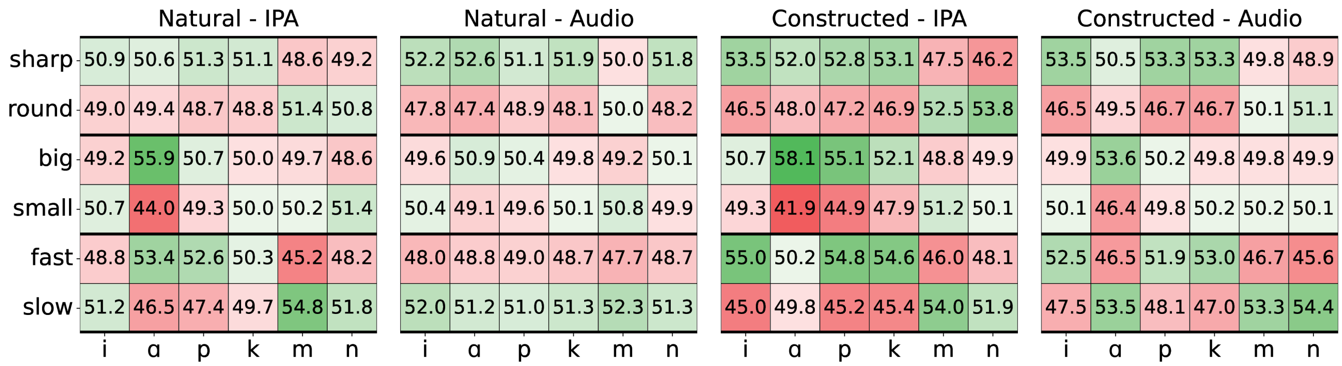

Figure 6: Attention fraction scores indicating the ratio of semantic dimensions that the model focuses on for each IPA, by word group and input type. For each IPA and semantic dimension, the model tends to attend more to semantic features that exhibit stronger associations with the given phonemes in 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).

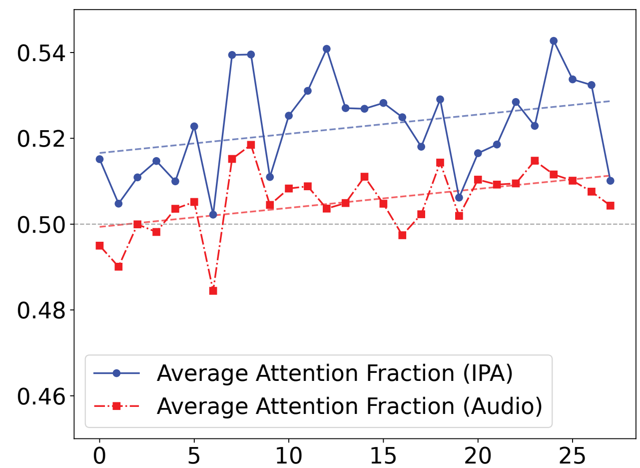

Figure 7: Attention fraction scores for the constructed word group, averaged across semantic dimensions. X-axis indicates the model's attention layer number, where Y-axis means the ratio at which the model attend to the correct semantic feature for a phoneme. While most layers score above the baseline 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.

Benczes, R. 2019. *Rhyme over reason: Phonological motivation in English*. Cambridge University Press.

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.

Centre national de la recherche scientifique; and Institut national de la langue française, eds. 1994. *Trésor de la langue française*, volume 16. Paris: CNRS Éditions, Gallimard. Dernier volume paru – tome XVI : Teint–Zzz.

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.

Enckell, P.; and Rézeau, P. 2003. *Dictionnaire des onomatopées*. Presses univ. de France.

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).

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Matsumura, A. 2019. Daijirin [Daijirin].

McAuliffe, M.; and Sonderegger, M. 2024a. English (US) ARPA acoustic model v3.0.0. Technical report, https://mfa-models.readthedocs.io/acoustic/English/English(US)ARPAacousticmodelv3_0_0.html.

McAuliffe, M.; and Sonderegger, M. 2024b. French MFA acoustic model v3.0.0. Technical report, https://mfa-models.readthedocs.io/acoustic/French/FrenchMFAacousticmodelv3_0_0.html.

McAuliffe, M.; and Sonderegger, M. 2024c. Japanese MFA acoustic model v3.0.0. Technical report, https://mfa-models.readthedocs.io/acoustic/Japanese/JapaneseMFAacousticmodelv3_0_0.html.

McAuliffe, M.; and Sonderegger, M. 2024d. Korean MFA acoustic model v3.0.0. Technical report, https://mfa-models.readthedocs.io/acoustic/Korean/KoreanMFAacousticmodelv3_0_0.html.

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. 2000. Standard Korean Language Dictionary Compilation Guideline 2. https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=31&pageIndex=51. Registered September 18, 2000; Last modified March 2, 2015; Accessed July 1, 2025.

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.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

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.

Shinmura, I. 2018. *Kojien*. Tokyo: Iwanami Shoten, 7 edition. ISBN 978-4000801318.

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.

Shogakukan. 2025. Digital Daijisen.

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.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

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.

Yamaguchi, N. 2003. *Kurashi no kotoba gion gitaigo jiten*. Kodansha.

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.

A Details on LEX-ICON

A.1 Mimetic Word Data

Natural Word Collection. We manually extract headwords and definitions from published onomatopoeia dictionaries for each language to determine whether a word belongs to the mimetic word category. For Korean, the word definition guideline of the National Institute of Korean Language (2000) is used to extract mimetic words and definitions. Semantic definitions that need to be supplemented are obtained from authoritative dictionaries for each language. The full list of the dictionaries is provided in Table 4.

We filter the data obtained from the mimetic word dictionaries, consisting of 954 English words, 1049 French words, 2025 Japanese words, and 4999 Korean words, using information from the aforementioned authoritative dictionaries. We finally determine 826 English words, 809 French words, 1418 Japanese words, and 4999 Korean words as natural mimetic word data.

Constructed Word Generation. We apply systematic pre-processing to IPA-dict entries from open-dict-data dictionaries during the filtering process to exclude words that have the same pronunciation as natural language words. We remove non-distinctive markers including stress (ˈ) and duration (ː) symbols, and convert allophones to their corresponding phonemes such as changing velarized lateral [ɫ] to /l/ in English and tap [ɾ] to /l/ in Korean. We also standardize notation by converting /eɪ/ to /ej/, /oʊ/ to /ow/.

For TTS compatibility, we create IPA-to-alphabet mapping rules and refine mappings where needed, which includes converting "tho" to "though" to ensure correct /ðow/ pronunciation. Generated words are hyphenated (e.g., "lah-mo") except for combinations ending in /ði/ (e.g., "laythey").

Word Group. In Figure 8, we illustrate example words of the *natural* and *constructed* word groups that comprise LEX-ICON.

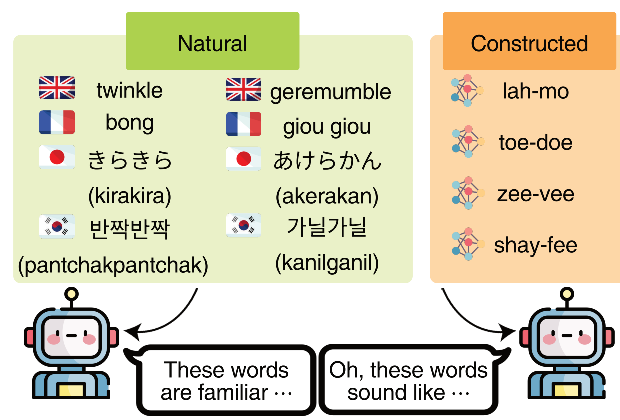

Figure 8: An illustrative description of the word groups. *Natural* word group consists of English, French, Japanese, and Korean mimetic words. *Constructed* word group includes disyllabic pseudo-words that experimental models may not have trained.

Input Type. We convert each word into three types (original orthographic text, phoneme-level spaced IPA text, and text-to-speech audio waveform) and input them into the experimental models. Figure 9 shows the semantic dimension

Language	References
English	Rhyme over Reason: Phonological Motivation in English (Benczes 2019) Oxford English Dictionary, 2nd Edition (Simpson 1989)
French	Dictionnaire des Onomatopées (Enckell and Rézeau 2003) Le Petit Robert de la Langue Française (Alain Rey 2022) Trésor de la Langue Française (Centre national de la recherche scientifique and Institut national de la langue française 1994)
Japanese	Kurashi no Kotoba Gion Gitaigo Jiten (Yamaguchi 2003) Seisenban Nihon Kokugo Daijiten (Shogakukan 2006) Kojien, 7th Edition (Shinmura 2018) Digital Daijisen (Shogakukan 2025) Daijirin, 4th Edition (Matsumura 2019)
Korean	Standard Korean Language Dictionary (National Institute of Korean Language 2025)

Table 4: Reference materials for natural mimetic word collection by language.

prediction performance of models for each word group and input type.

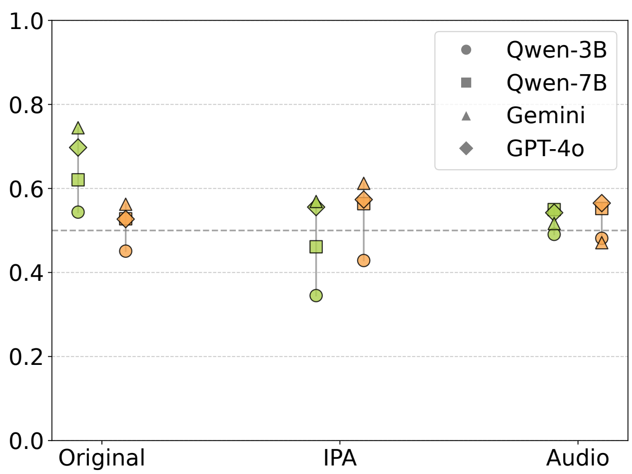

Figure 9: Semantic dimension prediction performance of models. For each input type, the dots for natural words' results are on the left, while constructed words' results are on the right. Each dot indicates macro-F1 scores of the models' response averaged across semantic dimensions.

A.2 Semantic Dimension

Table 6 shows the extreme feature pairs of 25 semantic dimensions suggested by Sidhu, Vigliocco, and Pexman (2022), which have been utilized for our semantic dimension prediction experiments.

After the automatic annotation process for natural words is complete, we utilize labels that are agreed upon by all model families and not annotated as "neither" as pseudo ground truth. Similarly, we remove semantic features close to neutral from the dataset by deleting semantic dimensions from each constructed word whose semantic features' ground truth score is within 1.0 standard deviation of the overall distribution. Due to this method, words in the datasets have fewer than 25 dimension labels. The distribution of these labels is shown in Table 5.

A.3 An Example Entry from LEX-ICON

LEX-ICON consists of the following information for each mimetic word, as shown in Table 7.

B Experimental Settings

B.1 Inference Settings

- Hardware: four NVIDIA RTX 4090 GPUs for 27B and 32B models, one NVIDIA RTX 4090 GPU for 3B and 7B models, and API calls for other proprietary models.
- Software: PyTorch (Paszke et al. 2017) and vLLM (Kwon et al. 2023) libraries.
- Random seed: 42 (for reproducibility, only if needed).
- Temperature: 0 (for reproducibility).
- Maximum output tokens: 1024.

B.2 Language Models Used in This Work

In this work, we utilize the following language models for data annotation, experimentation, and analysis.

- gpt-4o (OpenAI et al. 2024)
- gpt-4o-audio-preview (OpenAI et al. 2024)
- gpt-4.1 (OpenAI 2025)
- gemini-2.5-flash (Comanici et al. 2025)
- gemma-3-27b-it (Team et al. 2025b)
- Qwen3-32B (Yang et al. 2025a)
- Qwen2.5-Omni-3B (Xu et al. 2025)
- Qwen2.5-Omni-7B (Xu et al. 2025)

C Token-IPA Symbol Alignment

To enable phoneme-level attention analysis, we implement a pipeline for aligning audio tokens with IPA symbols.

Dimension	Natural			Constructed			Total		
(Feature 1)-(Feature 2)	Ft. 1	Ft. 2	Total	Ft. 1	Ft. 2	Total	Ft. 1	Ft. 2	Total
good-bad	532	1551	2083	0	0	0	532	1551	2083
beautiful-ugly	290	639	929	344	118	462	634	757	1391
pleasant-unpleasant	797	2583	3380	0	0	0	797	2583	3380
strong-weak	2273	2026	4299	107	101	208	2380	2127	4507
big-small	902	1171	2073	894	793	1687	1796	1964	3760
rugged-delicate	1327	1337	2664	0	0	0	1327	1337	2664
active-passive	3267	617	3884	0	0	0	3267	617	3884
fast-slow	1579	472	2051	750	687	1437	2329	1159	3488
sharp-round	1644	679	2323	789	834	1623	2433	1513	3946
realistic-fantastical	6876	7	6883	251	250	501	7127	257	7384
structured-disorganized	705	3007	3712	0	0	0	705	3007	3712
ordinary-unique	229	1356	1585	101	107	208	330	1463	1793
interesting-uninteresting	135	73	208	0	0	0	135	73	208
simple-complex	2064	2258	4322	784	818	1602	2848	3076	5924
abrupt-continuous	2055	2648	4703	440	565	1005	2495	3213	5708
exciting-calming	1741	515	2256	693	709	1402	2434	1224	3658
hard-soft	1757	1919	3676	440	696	1136	2197	2615	4812
happy-sad	436	283	719	709	754	1463	1145	1037	2182
harsh-mellow	2477	629	3106	440	565	1005	2917	1194	4111
heavy-light	1048	1870	2918	622	719	1341	1670	2589	4259
inhibited-free	1484	1189	2673	105	101	206	1589	1290	2879
masculine-feminine	144	234	378	689	833	1522	833	1067	1900
solid-nonsolid	933	1322	2255	105	788	893	1038	2110	3148
tense-relaxed	1986	970	2956	105	101	206	2091	1071	3162
dangerous-safe	392	56	448	440	101	541	832	157	989

Table 5: Semantic dimension distribution of pseudo ground truth by word group.

C.1 Montreal Forced Alignment (MFA) Pipeline

We employ Montreal Forced Alignment (McAuliffe et al. 2025) to obtain precise phoneme-level segmentation of audio waveforms. We utilize pre-trained acoustic models for each language:

- English (US) ARPA acoustic model v3.0.0 (McAuliffe and Sonderegger 2024a)
- French MFA acoustic model v3.0.0 (McAuliffe and Sonderegger 2024b)
- Japanese MFA acoustic model v3.0.0 (McAuliffe and Sonderegger 2024c)
- Korean MFA acoustic model v3.0.0 (McAuliffe and Sonderegger 2024d)

The alignment process generates Textgrid files containing phoneme boundaries with millisecond precision.

C.2 Custom Dictionary Generation and OOV Handling

We generate language-specific pronunciation dictionaries by mapping word-IPA pairs from our dataset to MFA phoneme representations. When initial alignment fails due to out-of-vocabulary (OOV) words, we automatically (1) identify words producing 'spn' (spoken noise) tokens in TextGrid output, (2) generate pronunciations using Epitran for OOV words, (3) update pronunciation dictionaries with new entries, and (4) re-run MFA alignment until convergence.

C.3 TextGrid Parsing and Audio Token Alignment

We parse MFA-generated TextGrid files to extract phoneme boundaries and convert them into discrete frame sequences compatible with the model's 40ms sampling period. For each audio frame, we assign the phoneme whose temporal boundary encompasses the frame's center point. The resulting frame-level sequence is then aligned with the model's audio token representations, enabling precise measurement of phoneme-level attention patterns for audio inputs.

D Metrics

D.1 Semantic Dimension Prediction

Semantic Dimension Macro-F1 Score.

- S: Set of semantic dimensions, $|S| = d$.
- C: Set of semantic features in each semantic dimension. $|C| = 2$.
- $TP_{s,c}$: True–positive count for dimension s and feature c.
- $FP_{s,c}$: False–positive count for dimension s and feature c.
- $FN_{s,c}$: False–negative count for dimension s and feature c.

The macro-F1 score for each semantic dimension s is defined as follows:

Feature A	Feature B
good	bad
beautiful	ugly
pleasant	unpleasant
strong	weak
big	small
rugged	delicate
active	passive
fast	slow
sharp	round
realistic	fantastical
structured	disorganized
ordinary	unique
interesting	uninteresting
simple	complex
abrupt	continuous
exciting	calming
hard	soft
happy	sad
harsh	mellow
heavy	light
inhibited	free
masculine	feminine
solid	nonsolid
tense	relaxed
dangerous	safe

Table 6: The 25 semantic dimension list by Sidhu, Vigliocco, and Pexman (2022).

$$\text{MacroF1}_s = \frac{1}{|C|}\sum_{c\in C}\frac{2TP_{s,c}}{2TP_{s,c}+FP_{s,c}+FN_{s,c}}$$

We calculate the final natural group scores for each semantic dimension of LLMs by averaging the macro-F1 scores for four natural languages (English, French, Japanese, and Korean) and three input types (original text, IPA text, and audio waveform) for each dimension. For the constructed words, only averaging three input types is performed with the same procedure as in the calculation of natural words' scores.

Correlation Score with Human Evaluation.

- D_{valid}: the set of semantic dimensions for which both human evaluation macro-F1 scores and a model's macro-F1 scores are available.
- h_d: the human evaluation macro-F1 score for dimension d and audio input type.
- m_d: the model's macro-F1 score for dimension d and a given input type.
- μ_H: the mean of the human scores across all dimensions in D_{valid}.

Key	Value
`word`	whizz
`meaning`	To make a sound as of a body rushing through the air.
`ref`	Rhyme Over Reason: Phonological Motivation in English
`ipa`	w ɪ z
`romanization`	whizz
`language`	en

(a) Metadata from an example word.

Dimension	Feature
`active-passive`	active
`fast-slow`	fast
`sharp-round`	sharp
`realistic-fantastical`	realistic
`simple-complex`	simple
`abrupt-continuous`	continuous
`exciting-calming`	exciting
`harsh-mellow`	harsh
`inhibited-free`	free
`solid-nonsolid`	nonsolid

(b) Semantic dimension data from an example word.

Table 7: An example entry from LEX-ICON.

- μ_M: the mean of the model's scores across all dimensions in D_{valid}.

The Pearson correlation coefficient r between the human scores (H) and the model's scores (M) for a given word group and input type is calculated as:

$$r(H,M) = \frac{\sum_{d\in D_{\text{valid}}}(h_d-\mu_H)(m_d-\mu_M)}{\sqrt{\sum_{d\in D_{\text{valid}}}(h_d-\mu_H)^2}\sqrt{\sum_{d\in D_{\text{valid}}}(m_d-\mu_M)^2}} \quad (1)$$

Audio Input Type Advantage Score. Let $S(g,t,d)$ be the macro F1-score for a given word group $g \in \{\text{natural, constructed}\}$, input type $t \in \{\text{audio, original text}\}$, and semantic dimension d.

The Input Type Advantage Score, denoted as $A(g,d)$, for a specific word group g and semantic dimension d is defined as the difference between the scores of the two input types:

$$A(g,d) = S(g,\text{audio},d) - S(g,\text{original text},d) \quad (2)$$

A positive value of $A(g,d)$ indicates that the "audio" input type has an advantage over the "original text" input type for that specific dimension and word group. Conversely, a negative value indicates an advantage for the "original text" input type.

The score $S(g,t,d)$ is calculated by averaging the macro F1-scores across all models for the given parameters:

$$S(g,t,d) = \frac{1}{|K|}\sum_{k\in K} s_{k,g,t,d} \quad (3)$$

Where:

- $s_{k,g,t,d}$: the macro F1-score of an individual model k.
- K: the set of all models evaluated.
- $|K|$: the total number of models.

D.2 Internal Attention Analysis

Attention Fraction Score Computation. The Phoneme-Semantic Feature Attention Fraction Score is calculated by the algorithm described in Algorithm 1.

E Human Evaluation

To ensure the reliability of the semantic dimension selection and the automatic annotation of the pseudo ground truth data in LEX-ICON, we conduct a human evaluation using 152 questions randomly sampled for each participant from the entire dataset. 10 graduate students without expertise in linguistics participate in this experiment, which is performed via the Label Studio (Tkachenko et al. 2020-2025). The experiment utilize questions with the same prompt structure as the semantic dimension prediction test that was provided to LLMs. With regard to the experiment's feasibility, all words are presented in audio waveforms directly hearable to humans. An example image of the process is displayed in Figure 10.

The randomly sampled test data contains 19 semantic dimensions that are presented in both of the natural and constructed word group data. For each dimension, all natural language types and answer semantic feature distributions are selected from the same distribution, respectively. As a result, four natural group words and four constructed group words are extracted for each dimension.

Due to the nature of sound symbolism task, in which disagreements may arise between responses, the human response distribution may have a wider variance than typical reasoning-centric experiments. However, we maximize reliability within limited resources through the aforementioned methodology. Table 8 shows human response average macro-F1 scores and standard deviation of the scores by semantic dimension.

Dimension	**Macro-F1 (± Std.)**
tense-relaxed	0.6784 (±0.2052)
realistic-fantastical	0.6333 (±0.2080)
masculine-feminine	0.6287 (±0.1643)
fast-slow	0.6210 (±0.2185)
big-small	0.6058 (±0.1298)
happy-sad	0.6003 (±0.1303)
abrupt-continuous	0.5903 (±0.2205)
beautiful-ugly	0.5844 (±0.1493)
heavy-light	0.5808 (±0.1648)
exciting-calming	0.5797 (±0.1948)
sharp-round	0.5796 (±0.1072)
hard-soft	0.5794 (±0.2142)
solid-nonsolid	0.5682 (±0.2259)
inhibited-free	0.5621 (±0.2263)
dangerous-safe	0.5594 (±0.1887)
harsh-mellow	0.5536 (±0.1652)
ordinary-unique	0.5376 (±0.1891)
strong-weak	0.4863 (±0.2000)
simple-complex	0.4602 (±0.1752)

Table 8: Human evaluation result with ranked average macro-F1 score and standard deviation by semantic dimension. Among the 25 semantic dimensions, 19 were used for human evaluation due to their presence in both natural and constructed word groups and their sufficient frequency in the dataset.

F Detailed Results

F.1 Semantic Dimension Prediction Experiment

Table 9, 10, 11, and 12 show accuracies and macro-F1 scores of experimental models' results for each input types and semantic dimensions. In the tables, "–" denotes a dimension where all ground truth features are classified as "neither" thus removed.

F.2 Internal Attention Analysis

We provide full result tables for the Phoneme-Semantic Feature Attention Fraction Scores by word group (natural and constructed) and input type (IPA text and audio waveform).

- Natural - IPA text: Table 13, 14, 15, and 16.
- Natural - Audio waveform: Table 17, 18, 19, and 20.
- Constructed - IPA text: Table 21 and 22.
- Constructed - Audio waveform: Table 23 and 24.

We also calculate the average attention fraction scores by layer in Qwen2.5-Omni-7B as shown in Table 25.

G Prompts

Table 26 shows an automatic semantic dimension annotation prompt for LLMs. Table 27 represents detailed prompts for the semantic dimension experiments.

Algorithm 1: Phoneme-Semantic Feature Attention Fraction Score Calculation.

```
Load model and prompt
model ← Qwen2.5-Omni-7B
data ← LOAD_DATA
ipa_list ← LOAD_IPA
dims ← LOAD_SEMANTIC_DIMENSIONS

Data format: {word: [(feature_1, feature_2), answer]}    ▷ feature_i: features from semantic dims

Model Inference Phase
for all word, (feature_1, feature_2, answer) in data do
    ipa_text ← TO_IPA(word)
    audio ← TO_AUDIO(word)
    for all input ∈ {ipa_text, audio} do
        prompt ← FORMAT_PROMPT(input, feature_1, feature_2)    ▷ Format prompt using input and semantic features
        response, attn, tokens ← INFERENCE(model, prompt)    ▷ attn = [layer, head, query, key]
        idx_input, idx_feature_1, idx_feature_2 ← FIND_IDX(tokens, input, feature_1, feature_2)    ▷ type: list[int]
        idxs ← {idx_input, idx_feature_1, idx_feature_2}    ▷ Aggregate relevant token indices
        filtered_attn ← attn[:, :, idxs, idxs]    ▷ Select sub-attention map for relevant tokens only
        SAVE(filtered_attn, tokens, input, idx_input, idx_feature_1, idx_feature_2, response)
    end for
end for

Phoneme-Semantic Feature Computation Phase
for all input ∈ {ipa_text, audio} do
    result ← {}
    for all word, (feature_1, feature_2, answer) in data do
        filtered_attn, tokens, input, idx_input, idx_feature_1, idx_feature_2, response ← LOAD(word, feature_1, feature_2)
        if answer ≠ response then
            continue
        end if
        ipa_span ← SPAN_ALIGNING_INPUT_TO_TOKENS(input, tokens, idx_input)
        for all layers ∈ filtered_attn do
            for all layer ∈ layers do    ▷ layer = [head, query, key]
                for all idx_phoneme ∈ ipa_span do
                    score_1 ← sum(layer[:, idx_feature_1, idx_phoneme])
                    score_2 ← sum(layer[:, idx_feature_2, idx_phoneme])
                    feature_1score ← score_1/(score_1 + score_2)
                    feature_2score ← score_2/(score_1 + score_2)
                    if result[phoneme][feature_1][layer] is undefined then
                        result[phoneme][feature_1][layer] ← []
                    end if
                    if result[phoneme][feature_2][layer] is undefined then
                        result[phoneme][feature_2][layer] ← []
                    end if
                    APPEND(result[phoneme][feature_1][layer], feature_1score)
                    APPEND(result[phoneme][feature_2][layer], feature_2score)
                end for
            end for
        end for
    end for
    for all phoneme ∈ ipa_list do
        for all feature ∈ dims do
            for all layer ∈ result[phoneme][feature] do
                result[phoneme][feature][layer] ← mean(result[phoneme][feature][layer])
            end for
        end for
    end for
    SAVE(result)
end for
```

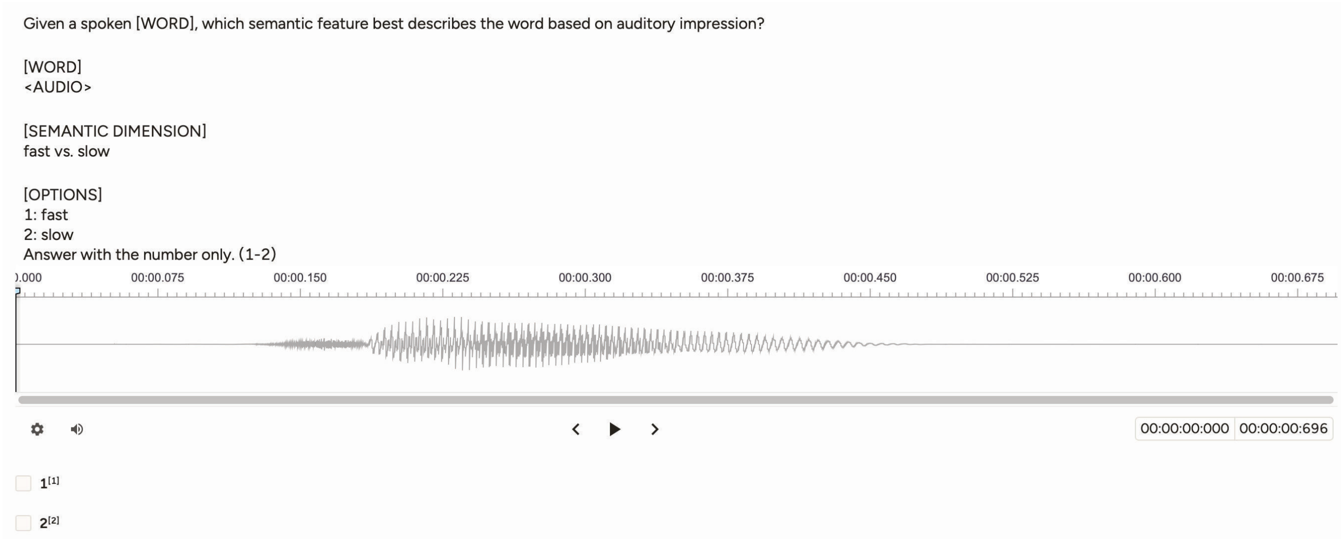

Figure 10: Screenshot image of the human evaluation process. Human evaluators are presented with the same prompts provided to LLMs via the Label Studio interface. Participants select a semantic feature after listening to audio waveform.

Dimension	**Natural**						**Constructed**					
	Original		IPA		Audio		Original		IPA		Audio	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
good-bad	67.0	61.9	38.9	38.4	53.1	48.8	–	–	–	–	–	–
beautiful-ugly	68.6	62.6	28.6	25.5	48.9	43.1	72.9	60.5	77.1	54.3	68.0	50.6
pleasant-unpleasant	68.4	63.6	43.9	43.0	56.5	53.1	–	–	–	–	–	–
strong-weak	71.0	68.2	55.8	52.1	66.4	62.6	60.1	60.0	52.4	51.9	65.9	64.3
big-small	61.6	54.3	62.5	60.9	60.5	54.1	49.1	39.0	64.4	64.0	50.3	46.9
rugged-delicate	65.9	64.8	50.1	48.7	59.9	58.3	–	–	–	–	–	–
active-passive	72.4	54.3	62.8	45.8	71.0	51.9	–	–	–	–	–	–
fast-slow	66.7	56.9	54.6	46.0	61.0	53.4	58.5	58.5	64.4	62.1	67.7	67.6
sharp-round	76.6	72.1	62.7	59.2	69.8	65.4	53.7	53.7	73.1	73.0	58.4	58.0
realistic-fantastical	68.0	39.8	19.7	15.9	75.3	43.0	60.3	53.4	58.5	51.8	79.8	79.8
structured-disorganized	69.8	65.0	24.7	21.5	57.4	54.8	–	–	–	–	–	–
ordinary-unique	76.8	50.8	74.3	48.0	68.2	48.9	51.9	35.9	64.4	58.2	63.9	61.0
interesting-uninteresting	72.9	63.2	66.0	38.7	67.6	45.6	–	–	–	–	–	–
simple-complex	69.0	44.1	70.8	49.1	68.2	41.2	49.8	34.6	49.0	33.0	48.8	33.2
abrupt-continuous	75.1	71.5	58.5	54.8	66.9	64.7	46.8	42.8	61.0	54.8	53.9	53.6
exciting-calming	51.3	47.3	22.1	19.9	42.9	39.9	52.6	41.3	59.9	52.0	53.6	48.4
hard-soft	71.7	70.9	55.2	52.8	64.7	63.1	57.4	53.5	64.8	58.7	57.2	55.8
happy-sad	75.8	74.4	54.4	48.4	62.2	58.2	70.3	70.0	56.7	51.1	69.0	68.1
harsh-mellow	84.9	68.8	55.8	48.7	78.5	61.2	54.7	51.3	67.3	66.8	51.5	46.7
heavy-light	74.7	73.7	56.1	55.3	66.9	65.7	64.8	64.5	50.4	47.8	46.1	43.2
inhibited-free	69.1	67.3	54.0	53.4	63.7	61.2	32.0	31.8	30.6	30.2	41.7	35.9
masculine-feminine	65.6	61.8	39.3	36.9	57.9	55.8	62.2	62.2	66.6	66.3	68.0	68.0
solid-nonsolid	75.8	69.1	63.1	55.5	70.7	61.1	87.5	52.9	77.4	56.2	81.1	60.6
tense-relaxed	74.7	66.8	56.5	50.9	64.6	56.6	55.8	55.8	77.7	77.5	65.5	64.6
dangerous-safe	70.5	58.7	54.8	43.0	53.6	45.3	61.6	49.4	57.9	50.0	59.0	52.7

Table 9: Detailed semantic dimension prediction accuracy and macro-F1 score results for Qwen2.5-Omni-7B.

Dimension	Natural						Constructed					
	Original		IPA		Audio		Original		IPA		Audio	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
good-bad	70.3	62.7	60.1	49.3	60.0	55.0	–	–	–	–	–	–
beautiful-ugly	72.6	64.6	64.0	47.3	60.9	56.1	43.7	40.5	38.7	38.7	67.5	52.1
pleasant-unpleasant	64.0	59.7	50.3	46.2	42.5	40.7	–	–	–	–	–	–
strong-weak	53.6	53.4	38.4	27.4	62.8	54.2	49.0	34.5	48.6	32.7	62.0	55.6
big-small	57.1	47.9	51.7	35.8	56.4	46.8	46.7	32.2	47.4	34.1	51.4	44.1
rugged-delicate	67.9	67.2	44.5	33.9	63.4	62.2	–	–	–	–	–	–
active-passive	49.9	42.8	14.4	12.5	76.2	53.0	–	–	–	–	–	–
fast-slow	51.1	47.2	26.1	23.8	53.2	47.0	58.9	55.6	49.1	39.6	61.7	61.7
sharp-round	77.2	69.4	63.3	59.4	80.3	70.4	55.5	53.8	73.2	72.7	54.3	47.2
realistic-fantastical	18.7	15.6	0.6	0.6	20.2	16.5	52.3	38.4	50.1	33.7	51.9	40.7
structured-disorganized	70.6	62.7	44.9	43.9	51.4	50.1	–	–	–	–	–	–
ordinary-unique	68.6	60.6	74.8	51.3	54.2	49.4	66.8	63.7	66.3	61.8	64.4	64.2
interesting-uninteresting	70.4	58.5	50.2	42.4	68.1	48.3	–	–	–	–	–	–
simple-complex	70.2	57.5	51.7	41.0	69.4	51.9	56.9	53.2	60.4	56.1	49.3	34.7
abrupt-continuous	69.2	66.8	46.2	34.7	63.7	60.2	42.7	34.0	55.9	35.9	54.7	48.8
exciting-calming	58.6	51.4	23.5	20.9	47.9	43.5	71.5	71.4	62.0	56.0	62.9	61.8
hard-soft	57.6	50.4	48.8	32.8	54.1	44.5	61.2	38.2	61.3	38.0	61.4	39.9
happy-sad	68.7	66.4	55.3	52.6	66.2	60.0	68.6	68.1	39.6	35.5	52.4	44.5
harsh-mellow	67.9	57.9	51.0	45.1	58.2	49.3	47.4	42.7	53.4	50.4	55.6	52.7
heavy-light	64.0	57.8	55.9	35.7	60.1	49.7	53.4	35.4	53.9	35.6	54.3	37.3
inhibited-free	61.6	54.4	53.0	34.4	56.9	43.7	47.6	33.0	49.0	32.9	48.1	32.5
masculine-feminine	71.0	64.0	44.2	32.2	66.9	60.2	58.0	49.8	56.9	41.3	71.5	70.7
solid-nonsolid	68.9	49.4	67.1	39.9	67.8	48.5	87.0	47.4	88.2	46.9	84.3	49.1
tense-relaxed	62.2	58.0	31.1	27.0	55.1	50.5	63.1	63.1	64.1	60.0	68.9	68.9
dangerous-safe	43.2	40.4	15.2	13.8	32.9	30.0	19.0	16.3	19.6	17.0	25.0	23.9

Table 10: Detailed semantic dimension prediction accuracy and macro-F1 score results for Qwen2.5-Omni-3B.

Dimension	Natural						Constructed					
	Original		IPA		Audio		Original		IPA		Audio	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
good-bad	76.1	71.4	48.9	46.8	54.2	51.3	–	–	–	–	–	–
beautiful-ugly	82.7	78.0	50.3	46.3	56.7	52.4	74.5	47.2	77.3	53.9	72.1	46.0
pleasant-unpleasant	75.8	71.3	50.4	48.0	51.5	49.8	–	–	–	–	–	–
strong-weak	78.8	77.5	65.7	61.7	64.3	63.3	77.4	77.4	67.8	67.7	77.9	77.9
big-small	83.4	82.1	69.4	68.4	68.8	65.9	63.0	59.3	75.3	74.8	64.5	62.2
rugged-delicate	80.7	79.8	65.7	64.0	67.7	65.7	–	–	–	–	–	–
active-passive	86.3	63.7	77.9	50.2	69.3	52.3	–	–	–	–	–	–
fast-slow	89.1	75.7	78.1	58.3	65.2	55.1	66.9	62.0	71.1	68.5	72.2	72.0
sharp-round	85.4	81.8	75.2	70.3	59.0	56.5	70.9	69.7	77.2	76.9	62.5	57.5
realistic-fantastical	23.9	18.8	13.4	10.7	32.5	23.6	50.5	35.3	39.3	30.7	63.1	61.9
structured-disorganized	83.9	76.7	37.2	36.8	52.3	51.1	–	–	–	–	–	–
ordinary-unique	77.4	51.0	77.6	58.4	68.3	46.8	51.0	33.8	61.1	53.3	51.9	43.1
interesting-uninteresting	70.9	50.6	68.8	52.7	70.4	58.5	–	–	–	–	–	–
simple-complex	73.1	58.2	74.0	57.3	71.5	55.5	52.7	41.0	49.9	35.1	54.6	46.3
abrupt-continuous	80.5	74.6	67.3	62.8	71.5	67.6	60.3	57.9	64.8	64.2	67.5	67.3
exciting-calming	84.1	73.5	55.9	50.8	49.8	45.4	66.2	65.9	74.0	73.3	54.4	46.7
hard-soft	80.5	80.2	64.5	63.9	67.2	65.2	63.4	45.4	71.6	64.1	71.2	64.2
happy-sad	79.4	74.7	58.8	52.1	60.1	58.0	49.4	35.5	51.9	40.2	62.5	61.5
harsh-mellow	75.5	65.5	53.1	47.0	47.8	43.5	59.8	47.6	60.6	46.4	62.2	52.7
heavy-light	81.6	79.6	70.6	69.3	68.1	61.7	57.4	43.6	73.9	71.0	55.6	40.8
inhibited-free	73.6	71.4	64.4	61.9	60.0	52.4	50.5	38.1	51.9	40.9	47.1	33.5
masculine-feminine	83.5	77.4	68.3	58.4	71.2	63.4	69.4	67.0	78.1	78.0	82.7	82.2
solid-nonsolid	81.0	76.8	63.0	60.8	69.5	66.3	87.9	60.2	67.7	59.3	70.8	61.4
tense-relaxed	79.4	73.4	66.3	57.6	68.0	61.8	71.8	71.7	68.0	66.3	85.9	85.9
dangerous-safe	63.2	54.9	45.0	39.9	28.8	26.0	40.9	40.6	27.7	27.1	19.4	16.7

Table 11: Detailed semantic dimension prediction accuracy and macro-F1 score results for gpt-4o.

Dimension	Natural						Constructed					
	Original		IPA		Audio		Original		IPA		Audio	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
good-bad	86.0	80.2	76.4	57.7	78.0	53.5	–	–	–	–	–	–
beautiful-ugly	90.7	86.7	72.2	60.1	62.4	55.7	72.7	65.1	67.1	64.3	39.4	38.0
pleasant-unpleasant	82.2	76.8	71.4	59.7	79.2	52.3	–	–	–	–	–	–
strong-weak	76.7	73.8	60.5	58.9	62.9	56.3	62.0	61.5	66.8	66.0	47.6	45.7
big-small	83.3	82.8	65.0	61.0	59.6	55.2	63.0	59.4	62.5	59.0	50.5	50.0
rugged-delicate	83.2	82.3	68.0	65.9	66.0	56.7	–	–	–	–	–	–
active-passive	87.5	59.8	70.2	49.6	68.2	49.6	–	–	–	–	–	–
fast-slow	80.7	70.0	47.6	44.1	30.9	30.5	63.9	63.9	58.0	51.9	50.9	41.5
sharp-round	86.5	82.2	73.3	68.8	76.9	65.3	61.2	60.2	73.1	71.9	56.6	52.1
realistic-fantastical	69.6	40.9	33.5	24.7	86.4	46.4	51.1	48.4	67.3	66.3	55.7	53.1
structured-disorganized	84.2	78.2	67.0	58.0	69.4	56.8	–	–	–	–	–	–
ordinary-unique	80.7	72.3	68.6	56.3	51.3	46.4	54.8	46.0	62.0	61.5	56.7	55.9
interesting-uninteresting	74.2	59.9	62.7	47.7	62.3	51.0	–	–	–	–	–	–
simple-complex	80.8	68.4	67.8	57.2	54.6	47.9	57.9	51.6	58.7	57.0	49.9	49.9
abrupt-continuous	84.2	81.3	73.0	69.6	71.7	61.5	55.4	55.3	61.7	57.5	50.0	48.3
exciting-calming	84.5	71.8	65.6	54.1	58.2	50.7	54.8	50.1	71.8	71.7	56.2	55.7
hard-soft	79.8	79.6	64.2	62.7	59.6	58.3	63.6	58.2	67.3	56.6	53.1	51.9
happy-sad	81.3	80.3	57.2	55.7	51.0	42.2	59.7	59.5	60.7	58.8	53.0	39.2
harsh-mellow	85.6	73.8	70.1	57.1	75.5	56.1	59.9	57.6	60.7	51.9	49.5	46.7
heavy-light	82.2	81.7	70.0	69.4	56.4	48.8	66.8	61.3	75.3	73.3	47.1	35.0
inhibited-free	74.4	73.9	60.9	60.1	55.8	51.6	46.6	44.7	47.1	44.4	50.0	40.9
masculine-feminine	82.5	76.2	59.6	52.7	54.0	51.7	67.0	61.8	65.0	60.5	58.1	58.0
solid-nonsolid	84.8	82.0	68.0	60.8	47.5	46.2	88.2	61.1	87.9	73.4	25.6	25.6
tense-relaxed	76.3	71.4	64.0	55.8	69.4	58.5	61.7	61.0	79.6	79.5	65.5	62.2
dangerous-safe	78.8	65.3	73.7	52.3	80.3	52.2	66.9	56.4	53.6	51.3	77.3	46.6

Table 12: Detailed semantic dimension prediction accuracy and macro-F1 score results for gemini-2.5-flash.

Dimension	i	ɪ	e	ɛ	æ	a	ɯ	u	ʊ	o	ʌ	ɔ	ɑ	ə
abrupt	.506	.491	.512	.511	.510	.517	.503	.506	.424	.509	.510	.492	.506	.511
continuous	.494	.509	.488	.489	.490	.482	.497	.494	.576	.490	.490	.508	.493	.489
active	.495	.499	.493	.482	.507	.502	.469	.496	.407	.496	.493	.461	.505	.492
passive	.505	.501	.507	.517	.493	.498	.530	.504	.593	.504	.507	.539	.495	.508
beautiful	.496	.455	.462	.499	.425	.487	.470	.443	.506	.484	.426	.442	.405	.486
ugly	.504	.545	.538	.501	.575	.513	.529	.556	.494	.516	.573	.558	.594	.514
big	.492	.473	.551	.452	.495	.469	.497	.515	.501	.465	.559	.531	.559	.506
small	.507	.527	.449	.547	.505	.531	.503	.485	.499	.535	.441	.469	.440	.494
dangerous	.507	.517	.512	.513	.528	.503	.505	.510	–	.507	.521	.505	.521	.510
safe	.493	.483	.488	.486	.471	.496	.495	.490	–	.493	.479	.494	.478	.490
exciting	.480	.484	.470	.478	.469	.471	.449	.474	–	.474	.468	.461	.442	.457
calming	.520	.516	.529	.521	.531	.527	.550	.526	–	.526	.531	.538	.558	.543
fast	.488	.570	.499	.521	.575	.519	.466	.478	–	.480	.492	.516	.534	.514
slow	.512	.429	.501	.479	.425	.481	.533	.522	–	.519	.507	.483	.465	.486
good	.500	.458	.468	.497	.417	.470	.487	.486	.591	.487	.475	.481	.453	.467
bad	.500	.542	.532	.503	.583	.529	.513	.514	.409	.512	.525	.519	.547	.533
happy	.522	.520	.484	.530	.471	.512	.508	.482	–	.469	.513	.517	.530	.502
sad	.478	.479	.516	.470	.528	.488	.492	.517	–	.530	.486	.482	.469	.498
hard	.479	.478	.476	.476	.484	.489	.473	.483	–	.478	.483	.470	.480	.475
soft	.521	.522	.524	.524	.516	.510	.526	.517	–	.522	.517	.530	.520	.525
harsh	.279	.266	.276	.262	.261	.283	.275	.283	.256	.273	.277	.275	.280	.258
mellow	.721	.734	.724	.738	.739	.717	.725	.717	.744	.727	.723	.725	.720	.741
heavy	.473	.459	.492	.466	.489	.471	.487	.500	.510	.478	.513	.520	.533	.482
light	.527	.541	.508	.534	.510	.529	.512	.500	.489	.522	.486	.480	.467	.518
inhibited	.512	.518	.478	.503	.481	.487	.504	.489	–	.504	.487	.491	.475	.490
free	.488	.482	.522	.497	.519	.513	.496	.511	–	.496	.513	.509	.525	.510
interesting	.520	.508	.491	.496	.492	.505	.503	.493	–	.506	.488	.492	.501	.497
uninteresting	.480	.492	.509	.503	.508	.495	.497	.507	–	.494	.512	.508	.499	.503
masculine	.495	.430	.483	.517	.433	.513	.483	.486	–	.512	.466	.477	.448	.480
feminine	.505	.569	.517	.483	.567	.487	.517	.514	–	.488	.534	.522	.552	.520
pleasant	.498	.491	.483	.505	.469	.497	.490	.478	.497	.493	.465	.496	.490	.499
unpleasant	.502	.509	.517	.494	.531	.503	.510	.521	.503	.507	.535	.504	.509	.501
realistic	.327	.341	.332	.314	.322	.318	.337	.320	.352	.316	.325	.312	.315	.342
fantastical	.673	.659	.668	.686	.678	.682	.663	.680	.647	.685	.675	.688	.685	.658
rugged	.474	.467	.477	.480	.453	.488	.489	.477	–	.484	.478	.469	.452	.463
delicate	.525	.532	.523	.519	.547	.512	.511	.523	–	.515	.522	.531	.547	.537
sharp	.509	.557	.483	.506	.582	.517	.459	.474	–	.478	.490	.474	.506	.503
round	.490	.442	.516	.494	.418	.482	.541	.525	–	.522	.509	.526	.494	.497
simple	.499	.520	.509	.511	.524	.503	.486	.497	.507	.507	.489	.511	.519	.517
complex	.501	.479	.491	.489	.476	.497	.513	.503	.493	.492	.511	.488	.481	.483
solid	.370	.351	.379	.359	.371	.370	.361	.366	–	.371	.364	.370	.363	.345
nonsolid	.629	.648	.621	.641	.629	.630	.639	.633	–	.628	.636	.629	.636	.655
strong	.483	.474	.482	.463	.501	.489	.473	.492	–	.481	.489	.522	.522	.480
weak	.517	.526	.517	.537	.499	.511	.527	.508	–	.519	.511	.478	.478	.520
structured	.361	.364	.346	.359	.360	.355	.359	.355	.376	.350	.350	.347	.360	.361
disorganized	.639	.636	.654	.641	.640	.645	.641	.645	.624	.650	.650	.653	.640	.639
tense	.501	.516	.494	.499	.515	.497	.496	.503	.494	.499	.499	.511	.497	.491
relaxed	.499	.484	.506	.501	.484	.502	.504	.497	.506	.501	.501	.489	.503	.509

Table 13: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - IPA text - Part 1).

Dimension	ũː	ɯː	aː	oː	iː	eː	b	d	g	p	t	k	v	ð
abrupt	.500	.502	.508	.504	.504	–	.523	.517	.501	.519	.512	.515	.500	.534
continuous	.500	.498	.492	.496	.496	–	.477	.482	.499	.481	.488	.485	.499	.466
active	.423	.465	.509	–	.492	.597	.497	.504	.491	.495	.498	.498	.482	–
passive	.577	.535	.491	–	.508	.403	.503	.496	.508	.505	.502	.502	.517	–
beautiful	–	.564	.432	–	.435	.404	.438	.419	.416	.483	.459	.444	.388	–
ugly	–	.435	.568	–	.565	.596	.561	.581	.584	.517	.540	.555	.612	–
big	–	.636	.596	.613	.533	–	.547	.567	.570	.507	.501	.500	.635	–
small	–	.364	.404	.387	.467	–	.452	.432	.430	.493	.499	.500	.365	–
dangerous	–	.539	.497	.495	.518	–	.532	.522	.529	.518	.524	.519	.531	–
safe	–	.461	.503	.505	.482	–	.468	.478	.471	.482	.475	.481	.468	–
exciting	.453	.464	–	–	.478	–	.467	.466	.479	.488	.484	.472	.437	.482
calming	.547	.536	–	–	.522	–	.533	.534	.520	.512	.516	.527	.562	.518
fast	–	.558	.574	–	.478	–	.519	.504	.517	.526	.514	.503	.542	.475
slow	–	.441	.426	–	.522	–	.480	.495	.483	.474	.485	.497	.458	.525
good	–	.486	.442	.439	.523	.427	.445	.451	.451	.493	.483	.477	.456	–
bad	–	.514	.558	.561	.476	.573	.555	.549	.549	.507	.517	.523	.544	–
happy	–	.458	.382	.422	.401	–	.459	.464	.474	.540	.511	.501	–	.424
sad	–	.542	.618	.578	.599	–	.541	.536	.525	.460	.488	.499	–	.576
hard	.473	.461	.453	–	.470	–	.489	.482	.492	.497	.491	.493	.464	.475
soft	.527	.538	.547	–	.530	–	.511	.517	.508	.503	.509	.507	.536	.525
harsh	–	.276	.292	.332	.258	.287	.300	.344	.274	.305	.305	.291	.317	–
mellow	–	.724	.708	.668	.742	.713	.700	.656	.726	.695	.695	.709	.683	–
heavy	.490	.546	.509	.514	.507	–	.539	.532	.530	.496	.499	.499	.627	–
light	.510	.454	.491	.486	.493	–	.461	.468	.470	.504	.501	.500	.373	–
inhibited	.458	.434	.454	–	.552	–	.483	.476	.509	.494	.500	.505	.459	–
free	.542	.566	.546	–	.448	–	.517	.524	.490	.506	.500	.495	.541	–
interesting	–	–	.457	–	–	–	.506	.496	.514	.516	.511	.503	–	–
uninteresting	–	–	.543	–	–	–	.494	.504	.486	.484	.489	.497	–	–
masculine	–	–	.408	.413	.469	–	.455	.468	.492	.536	.512	.520	–	–
feminine	–	–	.591	.587	.531	–	.545	.532	.508	.464	.488	.480	–	–
pleasant	–	.497	.446	.438	.455	.430	.480	.477	.482	.488	.480	.471	.450	–
unpleasant	–	.503	.554	.562	.545	.570	.520	.523	.517	.512	.520	.529	.550	–
realistic	.322	.322	.304	.285	.308	.357	.328	.317	.308	.335	.326	.335	.259	–
fantastical	.678	.678	.696	.715	.692	.643	.672	.683	.692	.665	.674	.666	.741	–
rugged	.468	.459	.498	.471	.479	–	.488	.471	.488	.494	.489	.490	–	.483
delicate	.532	.541	.502	.529	.521	–	.512	.529	.512	.505	.511	.510	–	.517
sharp	–	.460	–	.486	.523	–	.504	.499	.499	.513	.512	.511	.519	–
round	–	.540	–	.514	.477	–	.496	.501	.501	.487	.488	.488	.480	–
simple	.511	.522	.528	.545	.503	.523	.520	.512	.499	.510	.489	.484	.515	.519
complex	.489	.477	.472	.455	.497	.476	.479	.487	.501	.490	.510	.516	.485	.481
solid	.346	.376	.370	.361	.382	–	.385	.358	.386	.380	.374	.371	.390	–
nonsolid	.654	.624	.630	.639	.618	–	.615	.642	.614	.620	.626	.629	.610	–
strong	.461	.538	.522	.510	.552	.617	.525	.533	.523	.500	.497	.495	.565	–
weak	.539	.462	.478	.490	.448	.383	.475	.467	.477	.500	.503	.504	.435	–
structured	.417	.386	–	.354	.372	–	.337	.326	.347	.353	.363	.358	.335	–
disorganized	.583	.614	–	.646	.628	–	.663	.674	.653	.647	.637	.642	.665	–
tense	.485	.552	.529	–	.543	–	.524	.502	.521	.515	.518	.506	.513	–
relaxed	.515	.448	.471	–	.457	–	.476	.498	.479	.485	.481	.494	.487	–

Table 14: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - IPA text - Part 2).

Dimension	z	ʒ	f	s	ʃ	θ	ç	h	ɸ	m	n	ɲ	ŋ	N
abrupt	.515	.488	.528	.508	.516	.510	.505	.516	.515	.511	.511	.504	.500	.474
continuous	.485	.512	.472	.491	.484	.490	.494	.484	.485	.489	.489	.496	.500	.526
active	.503	.495	.510	.494	.505	.513	.496	.507	.495	.497	.488	.483	.493	.468
passive	.497	.505	.490	.506	.495	.487	.504	.492	.504	.502	.511	.517	.507	.532
beautiful	.451	.396	.435	.490	.451	.409	.491	.474	.418	.455	.453	.485	.487	.501
ugly	.548	.604	.565	.510	.549	.591	.509	.526	.582	.545	.547	.515	.513	.499
big	.507	.559	.499	.473	.561	.617	.459	.510	.491	.497	.486	–	.507	.535
small	.493	.441	.501	.527	.439	.383	.541	.490	.509	.502	.514	–	.493	.465
dangerous	.532	.542	.511	.528	.529	.522	.557	.521	.516	.526	.517	.550	.515	.514
safe	.467	.458	.489	.472	.471	.477	.442	.479	.484	.474	.482	.450	.485	.486
exciting	.494	.526	.501	.486	.507	–	.482	.487	.486	.474	.473	.491	.466	.446
calming	.505	.474	.498	.513	.493	–	.517	.512	.514	.526	.527	.509	.533	.554
fast	.504	.485	.556	.483	.563	.520	.523	.514	.513	.452	.482	–	.476	.491
slow	.496	.515	.444	.516	.437	.480	.477	.486	.487	.548	.518	–	.523	.509
good	.485	.410	.472	.493	.476	–	.484	.489	.465	.476	.481	.481	.495	.468
bad	.515	.590	.528	.507	.524	–	.516	.510	.535	.523	.519	.519	.505	.532
happy	.437	–	.508	.516	.472	.484	.410	.529	.563	.492	.488	.549	.514	.436
sad	.563	–	.492	.484	.528	.516	.590	.470	.437	.508	.511	.451	.486	.563
hard	.471	.475	.484	.482	.480	.483	.483	.504	.492	.486	.493	.484	.483	.476
soft	.528	.525	.515	.518	.520	.517	.517	.496	.508	.513	.506	.516	.517	.524
harsh	.321	.264	.305	.291	.295	.283	.304	.321	.262	.269	.286	.263	.276	.268
mellow	.679	.736	.695	.709	.705	.717	.696	.679	.738	.731	.714	.737	.724	.732
heavy	.523	.524	.497	.478	.514	.545	.483	.469	.496	.500	.479	.460	.509	.512
light	.477	.476	.503	.522	.486	.455	.517	.531	.504	.500	.521	.540	.491	.488
inhibited	.478	.500	.481	.504	.448	.577	.498	.507	.455	.523	.504	.494	.497	.483
free	.522	.500	.519	.496	.552	.423	.502	.493	.545	.477	.496	.506	.503	.517
interesting	.510	–	.506	.507	.507	–	.512	.504	–	.498	.475	.513	.487	–
uninteresting	.490	–	.494	.492	.493	–	.488	.496	–	.501	.525	.487	.513	–
masculine	.414	.459	.475	.533	–	–	–	.491	.491	.484	.478	.530	.523	.484
feminine	.586	.540	.525	.467	–	–	–	.509	.509	.515	.522	.470	.477	.516
pleasant	.486	.430	.471	.493	.475	.459	.490	.491	.492	.478	.476	.480	.488	.490
unpleasant	.513	.570	.529	.507	.525	.541	.510	.509	.507	.522	.524	.520	.512	.510
realistic	.297	.309	.337	.321	.320	.322	.297	.331	.284	.336	.337	.321	.327	.361
fantastical	.703	.691	.663	.679	.680	.678	.703	.669	.716	.664	.663	.679	.673	.639
rugged	.493	.452	.478	.496	.461	.445	.478	.505	.491	.486	.494	.496	.503	.500
delicate	.507	.548	.521	.503	.538	.554	.522	.495	.509	.514	.506	.504	.497	.500
sharp	.531	.467	.540	.511	.564	.536	.518	.517	.515	.486	.492	.492	.481	.477
round	.469	.533	.460	.489	.436	.463	.482	.483	.485	.514	.508	.508	.519	.523
simple	.504	.490	.515	.503	.513	.518	.501	.513	.512	.494	.492	.486	.490	.490
complex	.495	.510	.485	.496	.487	.482	.499	.487	.488	.506	.508	.514	.510	.509
solid	.373	.365	.368	.350	.363	.380	.370	.372	.387	.360	.366	.355	.350	.349
nonsolid	.627	.635	.632	.650	.637	.619	.630	.627	.613	.640	.634	.644	.650	.651
strong	.534	.523	.506	.479	.538	.525	.488	.494	.478	.477	.487	.472	.496	.497
weak	.466	.477	.494	.521	.462	.475	.512	.506	.522	.523	.513	.527	.504	.502
structured	.327	.366	.323	.340	.364	.360	.351	.316	.306	.338	.333	.361	.354	.360
disorganized	.673	.634	.677	.660	.635	.639	.649	.684	.694	.662	.667	.639	.646	.640
tense	.523	.551	.536	.501	.540	.542	.515	.499	.512	.504	.498	.514	.499	.487
relaxed	.476	.449	.464	.498	.460	.458	.485	.501	.488	.496	.502	.485	.500	.513

Table 15: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - IPA text - Part 3).

Dimension	l	R	ɾ	ɹ	j	ɰ	ɕ	d͡ʑ	t͡ɕ	ʑ	ʔ	w	ɥ
abrupt	.507	.509	.506	.502	.518	.493	.514	.515	.511	.500	.534	.507	.511
continuous	.492	.490	.494	.497	.482	.507	.485	.485	.489	.500	.466	.492	.489
active	.491	.478	.498	.495	.512	.511	.508	.492	.503	.501	.475	.498	.471
passive	.508	.521	.501	.505	.488	.489	.492	.508	.497	.499	.525	.501	.529
beautiful	.463	.432	.496	.448	.460	.514	.447	.368	.441	.425	.455	.443	.411
ugly	.537	.568	.504	.551	.540	.486	.552	.632	.559	.574	.545	.557	.589
big	.484	.554	.508	.506	.468	.587	.505	.527	.503	.480	.520	.537	.491
small	.516	.445	.492	.494	.532	.413	.495	.473	.497	.520	.480	.463	.508
dangerous	.513	.527	.522	.548	.504	–	.521	.519	.518	.509	.506	.514	.509
safe	.486	.473	.477	.451	.496	–	.478	.481	.482	.491	.494	.486	.491
exciting	.465	.474	.482	.481	.493	.445	.486	.488	.481	.477	.466	.470	–
calming	.535	.526	.517	.518	.507	.555	.514	.512	.519	.523	.533	.530	–
fast	.482	.504	.468	.519	.485	.452	.506	.474	.486	.465	.534	.528	.538
slow	.518	.496	.531	.481	.515	.548	.494	.526	.514	.534	.465	.472	.462
good	.474	.472	.468	.462	.478	.473	.489	.482	.478	.439	.468	.476	.490
bad	.526	.528	.532	.538	.522	.527	.511	.518	.522	.561	.532	.524	.510
happy	.500	.528	.492	.491	.482	.476	.476	.446	.453	.385	.482	.506	–
sad	.499	.472	.508	.509	.518	.524	.523	.554	.547	.614	.518	.494	–
hard	.482	.489	.494	.485	.485	.466	.497	.487	.505	.469	.487	.484	.487
soft	.518	.511	.505	.515	.515	.534	.503	.513	.495	.531	.513	.516	.513
harsh	.287	.275	.281	.278	.296	.277	.297	.323	.284	.296	.275	.298	.285
mellow	.713	.725	.719	.722	.704	.723	.703	.677	.716	.704	.725	.702	.715
heavy	.475	.528	.487	.496	.470	.470	.500	.517	.503	.510	.492	.516	.488
light	.525	.472	.512	.504	.530	.530	.499	.483	.497	.490	.508	.484	.512
inhibited	.492	.469	.471	.481	.484	.472	.510	.501	.505	.525	.485	.476	.462
free	.508	.530	.529	.519	.515	.527	.490	.499	.495	.475	.515	.524	.538
interesting	.507	.484	.513	.504	.503	–	.528	–	.512	–	.493	.509	–
uninteresting	.493	.516	.486	.496	.497	–	.471	–	.488	–	.507	.491	–
masculine	.501	–	.487	.447	.498	.489	.466	–	.503	.481	.477	.499	–
feminine	.499	–	.513	.553	.502	.511	.534	–	.497	.519	.523	.500	–
pleasant	.482	.494	.499	.482	.484	.506	.476	.465	.465	.458	.483	.484	.513
unpleasant	.517	.505	.500	.518	.516	.494	.523	.535	.535	.542	.517	.516	.486
realistic	.336	.321	.331	.329	.310	.323	.288	.262	.295	.332	.339	.330	.355
fantastical	.665	.679	.668	.671	.690	.677	.712	.738	.705	.668	.661	.670	.645
rugged	.487	.494	.509	.478	.487	.489	.487	.508	.486	.506	.503	.486	.481
delicate	.513	.506	.490	.522	.512	.510	.513	.492	.514	.493	.497	.514	.518
sharp	.481	.500	.478	.501	.525	.466	.533	.495	.528	.491	.537	.525	.550
round	.519	.500	.522	.499	.475	.534	.466	.505	.472	.509	.463	.474	.449
simple	.495	.499	.495	.508	.500	.510	.462	.496	.471	.439	.502	.511	.536
complex	.505	.501	.505	.492	.499	.490	.537	.504	.529	.560	.498	.488	.464
solid	.367	.367	.369	.354	.353	.372	.367	.387	.374	.365	.385	.370	.350
nonsolid	.633	.632	.631	.646	.647	.628	.633	.613	.626	.635	.615	.630	.650
strong	.487	.525	.492	.489	.497	.472	.512	.516	.504	.480	.502	.504	.469
weak	.513	.475	.508	.511	.503	.528	.488	.484	.496	.520	.498	.496	.531
structured	.338	.366	.364	.350	.346	.356	.383	.330	.382	.350	.342	.338	.360
disorganized	.662	.634	.636	.650	.654	.644	.617	.670	.618	.649	.658	.662	.640
tense	.497	.509	.496	.507	.511	.488	.545	.509	.542	.526	.519	.513	.517
relaxed	.503	.491	.504	.493	.489	.512	.454	.491	.458	.474	.481	.487	.482

Table 16: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - IPA text - Part 4).

Dimension	i	ɪ	e	ɛ	a	ɨ	ɯ	u	ʊ	o	ɚ	ʌ	ɔ	ɑ
abrupt	.517	.525	.526	.522	.537	.518	.513	.513	.557	.516	.534	.524	.520	.532
continuous	.483	.475	.474	.478	.463	.482	.486	.487	.443	.484	.466	.476	.480	.468
active	.500	.506	.496	.496	.502	.494	.504	.496	.485	.497	.509	.490	.511	.511
passive	.500	.494	.504	.504	.498	.506	.496	.504	.515	.503	.491	.510	.489	.489
beautiful	.467	.467	.460	.488	.479	.469	.479	.439	.531	.470	.479	.437	.475	.466
ugly	.533	.533	.540	.512	.521	.531	.521	.561	.469	.530	.521	.563	.525	.534
big	.496	.511	.513	.488	.505	.494	.498	.496	.525	.484	.516	.508	.517	.509
small	.504	.489	.487	.512	.495	.506	.502	.504	.475	.516	.484	.492	.483	.491
dangerous	.520	.526	.533	.526	.524	.532	.510	.519	–	.518	.517	.512	.526	.516
safe	.480	.474	.467	.474	.476	.468	.490	.481	–	.482	.483	.488	.474	.484
exciting	.497	.515	.499	.507	.487	.491	.490	.494	–	.489	.503	.493	.505	.499
calming	.502	.485	.501	.492	.513	.509	.510	.506	–	.511	.497	.507	.495	.501
fast	.480	.495	.492	.486	.486	.476	.482	.468	–	.463	.492	.477	.494	.488
slow	.520	.505	.508	.514	.514	.524	.518	.532	–	.537	.508	.523	.506	.512
good	.498	.500	.477	.487	.486	.501	.499	.499	.562	.499	.505	.491	.496	.491
bad	.502	.500	.523	.513	.514	.499	.501	.501	.438	.501	.495	.509	.504	.509
happy	.487	.467	.491	.487	.488	.488	.475	.486	–	.484	.487	.480	.484	.482
sad	.513	.533	.509	.512	.512	.512	.525	.514	–	.516	.513	.520	.516	.518
hard	.495	.499	.503	.505	.501	.497	.487	.493	–	.494	.487	.499	.495	.500
soft	.504	.501	.497	.494	.499	.503	.513	.507	–	.506	.513	.500	.505	.500
harsh	.299	.315	.309	.320	.294	.315	.286	.314	.319	.301	.308	.329	.306	.311
mellow	.701	.685	.691	.680	.706	.685	.714	.686	.681	.699	.692	.671	.694	.689
heavy	.511	.504	.519	.506	.520	.510	.516	.521	.528	.517	.515	.526	.530	.521
light	.489	.496	.481	.494	.480	.490	.484	.479	.472	.483	.485	.474	.470	.479
inhibited	.504	.491	.502	.492	.510	.500	.507	.493	–	.501	.496	.499	.489	.492
free	.496	.509	.498	.508	.490	.500	.493	.507	–	.499	.504	.501	.511	.508
interesting	.504	.500	.519	.510	.507	.498	.496	.510	–	.507	–	.503	.511	.504
uninteresting	.496	.500	.481	.490	.493	.502	.504	.490	–	.493	–	.497	.489	.496
masculine	.501	.486	.494	.529	.496	.517	.496	.506	–	.503	.493	.513	.494	.497
feminine	.498	.514	.506	.471	.504	.483	.504	.494	–	.497	.507	.487	.506	.503
pleasant	.492	.502	.481	.494	.493	.487	.492	.482	.503	.489	.504	.482	.494	.497
unpleasant	.508	.498	.519	.506	.507	.513	.508	.518	.497	.511	.496	.518	.506	.503
realistic	.353	.339	.352	.357	.333	.341	.336	.342	.365	.342	.353	.342	.333	.335
fantastical	.647	.661	.648	.643	.667	.659	.664	.658	.635	.658	.647	.658	.667	.665
rugged	.465	.485	.470	.473	.474	.475	.474	.479	–	.478	.488	.479	.485	.485
delicate	.535	.515	.530	.527	.526	.525	.526	.521	–	.522	.512	.521	.515	.515
sharp	.522	.533	.515	.528	.522	.513	.506	.498	–	.499	.523	.523	.519	.526
round	.478	.467	.485	.472	.478	.487	.494	.502	–	.501	.476	.477	.481	.474
simple	.513	.514	.508	.506	.510	.507	.509	.506	.525	.507	.510	.494	.508	.513
complex	.487	.486	.492	.494	.490	.493	.491	.494	.475	.493	.490	.506	.492	.487
solid	.376	.362	.376	.384	.375	.365	.375	.378	–	.382	.362	.379	.380	.370
nonsolid	.624	.638	.624	.616	.625	.635	.625	.622	–	.618	.638	.621	.620	.630
strong	.500	.496	.497	.504	.501	.500	.498	.502	–	.503	.502	.505	.510	.511
weak	.500	.504	.503	.496	.499	.500	.502	.498	–	.497	.498	.495	.490	.489
structured	.343	.350	.343	.355	.339	.346	.348	.354	.369	.346	.336	.342	.345	.343
disorganized	.657	.650	.657	.645	.661	.654	.652	.646	.631	.654	.664	.658	.655	.657
tense	.512	.512	.516	.517	.503	.517	.502	.512	.520	.511	.492	.520	.507	.504
relaxed	.488	.488	.484	.483	.497	.483	.498	.487	.480	.489	.508	.480	.493	.496

Table 17: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - Audio - Part 1).

Dimension	ɐ	ə	ɯː	aː	oː	iː	b	d	g	p	t	c	k	v
abrupt	.530	.536	.515	.541	.546	.555	.516	.518	.515	.518	.524	.514	.521	.505
continuous	.470	.464	.485	.459	.454	.445	.484	.482	.485	.482	.476	.486	.479	.495
active	.494	.503	.508	.516	.508	.504	.507	.512	.506	.504	.497	.508	.502	.509
passive	.506	.497	.492	.484	.492	.496	.493	.488	.494	.496	.503	.492	.498	.491
beautiful	.464	.471	.488	–	–	.444	.469	.455	.479	.491	.469	.505	.463	.430
ugly	.536	.529	.512	–	–	.556	.531	.545	.521	.509	.531	.495	.537	.570
big	.486	.505	–	.468	.513	–	.512	.520	.505	.504	.501	.494	.498	.501
small	.514	.495	–	.532	.487	–	.488	.480	.495	.496	.499	.506	.501	.499
dangerous	.512	.526	.487	.532	.543	.468	.517	.521	.516	.497	.522	.512	.518	.503
safe	.488	.473	.513	.468	.457	.532	.483	.479	.484	.503	.478	.488	.482	.497
exciting	.487	.488	.514	.484	.430	.508	.489	.485	.493	.496	.490	.505	.498	.496
calming	.512	.512	.486	.516	.570	.492	.510	.515	.507	.504	.510	.495	.502	.504
fast	.479	.477	.506	.482	–	–	.490	.483	.484	.490	.482	.480	.487	.514
slow	.521	.523	.494	.518	–	–	.510	.517	.516	.510	.517	.520	.513	.486
good	.480	.493	.514	.499	.479	.454	.487	.489	.504	.499	.495	.507	.496	.455
bad	.520	.507	.486	.501	.521	.546	.513	.511	.496	.501	.505	.493	.504	.545
happy	.479	.481	.496	.492	–	–	.480	.522	.488	.499	.488	.475	.481	–
sad	.521	.518	.504	.508	–	–	.520	.478	.512	.501	.512	.525	.519	–
hard	.507	.500	.478	.504	.498	.503	.488	.492	.493	.494	.499	.482	.492	.491
soft	.493	.500	.522	.496	.502	.497	.512	.508	.507	.506	.501	.518	.508	.509
harsh	.324	.317	.298	.298	.275	.298	.285	.286	.292	.298	.315	.297	.307	.276
mellow	.676	.683	.703	.702	.725	.702	.715	.714	.708	.702	.685	.703	.693	.724
heavy	.509	.521	.515	.531	.535	.493	.523	.520	.515	.517	.521	.510	.515	.528
light	.491	.479	.485	.469	.465	.507	.477	.480	.485	.483	.479	.490	.485	.472
inhibited	.501	.500	.504	.503	.505	.485	.505	.507	.499	.494	.502	.502	.502	.495
free	.499	.500	.496	.497	.495	.515	.495	.492	.501	.506	.498	.497	.498	.505
interesting	.507	.503	–	.522	–	–	.503	.491	.496	.504	.504	.525	.504	–
uninteresting	.493	.497	–	.478	–	–	.497	.509	.504	.496	.496	.475	.496	–
masculine	.519	.461	–	.514	.508	–	.501	.500	.504	.505	.507	.482	.516	–
feminine	.481	.539	–	.486	.492	–	.499	.500	.496	.495	.493	.518	.484	–
pleasant	.488	.496	.506	.461	.493	.483	.497	.496	.498	.498	.492	.513	.491	.497
unpleasant	.512	.504	.494	.539	.507	.517	.503	.504	.502	.502	.508	.487	.508	.503
realistic	.346	.332	.350	.343	.333	.350	.352	.356	.358	.358	.347	.342	.353	.348
fantastical	.655	.668	.650	.657	.667	.650	.648	.644	.642	.642	.653	.658	.648	.652
rugged	.478	.489	.466	.475	.482	.457	.471	.459	.477	.479	.476	.474	.476	.473
delicate	.522	.511	.534	.525	.518	.543	.529	.541	.523	.521	.524	.526	.524	.527
sharp	.528	.512	.517	.520	–	.539	.504	.507	.510	.511	.518	.535	.519	.521
round	.471	.487	.483	.480	–	.461	.496	.493	.490	.489	.482	.465	.481	.479
simple	.505	.511	.521	.521	.516	.517	.520	.518	.517	.519	.511	.514	.511	.517
complex	.495	.489	.479	.479	.484	.483	.480	.482	.483	.481	.489	.486	.489	.483
solid	.382	.374	.371	.375	.390	.364	.371	.371	.370	.372	.378	.374	.377	.416
nonsolid	.618	.626	.629	.625	.610	.636	.629	.629	.630	.628	.621	.626	.623	.584
strong	.506	.507	.510	.512	.493	.514	.504	.502	.506	.502	.506	.501	.500	.508
weak	.494	.493	.490	.488	.507	.486	.496	.498	.494	.498	.494	.499	.500	.492
structured	.349	.341	.347	.351	.318	.354	.339	.344	.344	.347	.349	.351	.344	.335
disorganized	.651	.659	.653	.649	.682	.646	.661	.656	.656	.653	.651	.649	.656	.665
tense	.511	.512	.503	.509	.515	.506	.496	.503	.495	.502	.513	.511	.506	.492
relaxed	.489	.488	.497	.491	.485	.494	.504	.497	.505	.498	.487	.489	.494	.508

Table 18: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - Audio - Part 2).

Dimension	z	ʁ	f	s	ʃ	θ	h	ɸ	m	n	ɲ	ŋ	N	l
abrupt	.514	.534	.514	.505	.525	.521	.514	.513	.511	.512	.500	.519	.522	.518
continuous	.485	.466	.486	.495	.475	.479	.486	.487	.489	.488	.500	.481	.477	.482
active	.508	.502	.506	.499	.508	.516	.505	.503	.505	.505	.502	.500	.520	.505
passive	.492	.498	.494	.501	.492	.484	.495	.497	.495	.495	.498	.500	.480	.495
beautiful	.464	.442	.498	.487	.495	.474	.488	.529	.472	.473	.491	.482	.505	.484
ugly	.535	.558	.502	.513	.504	.526	.512	.471	.528	.527	.509	.518	.495	.516
big	.484	.506	.509	.484	.512	.478	.504	.498	.492	.501	–	.495	.500	.506
small	.516	.494	.490	.516	.488	.521	.495	.502	.508	.499	–	.505	.500	.494
dangerous	.517	.536	.506	.499	.508	.502	.506	.482	.515	.525	.501	.518	.521	.509
safe	.483	.464	.494	.501	.492	.498	.494	.517	.485	.475	.499	.482	.478	.491
exciting	.500	.495	.497	.498	.495	.499	.498	.494	.493	.497	.496	.501	.496	.496
calming	.500	.505	.503	.502	.505	.501	.502	.506	.507	.503	.504	.499	.504	.504
fast	.484	.495	.501	.477	.508	.482	.489	.462	.477	.487	.471	.485	.496	.479
slow	.516	.505	.499	.523	.492	.518	.511	.538	.523	.513	.529	.515	.504	.521
good	.487	.474	.499	.503	.493	–	.505	.491	.500	.504	.518	.502	.508	.494
bad	.512	.526	.501	.497	.507	–	.495	.509	.500	.496	.482	.498	.492	.506
happy	.484	.490	.472	.484	.478	.478	.502	.499	.490	.484	.496	.489	.493	.489
sad	.516	.510	.528	.515	.522	.522	.497	.501	.510	.516	.504	.511	.507	.511
hard	.489	.499	.485	.486	.486	.491	.493	.472	.494	.497	.485	.497	.489	.493
soft	.511	.501	.515	.514	.514	.509	.507	.527	.506	.503	.515	.503	.510	.507
harsh	.298	.290	.298	.305	.301	.280	.298	.291	.294	.288	.277	.301	.289	.303
mellow	.702	.710	.702	.695	.699	.720	.702	.709	.706	.712	.723	.699	.711	.697
heavy	.503	.532	.514	.508	.517	.510	.506	.498	.524	.520	.496	.518	.515	.517
light	.497	.468	.486	.492	.483	.490	.494	.502	.476	.480	.504	.482	.485	.483
inhibited	.504	.474	.493	.498	.492	.511	.498	.499	.500	.505	.506	.506	.501	.499
free	.495	.526	.507	.502	.508	.489	.502	.500	.500	.495	.494	.494	.499	.501
interesting	.503	.514	.514	.504	.519	–	.506	–	.501	.498	.501	.495	.518	.500
uninteresting	.497	.486	.486	.496	.481	–	.494	–	.499	.502	.499	.505	.482	.500
masculine	–	.466	.498	.512	.515	–	.490	.475	.504	.500	.506	.518	.501	.509
feminine	–	.534	.502	.488	.485	–	.510	.525	.496	.500	.494	.481	.498	.491
pleasant	.489	.495	.499	.497	.494	.510	.495	.490	.488	.492	.503	.492	.500	.501
unpleasant	.511	.505	.501	.503	.506	.489	.505	.510	.512	.508	.497	.508	.500	.499
realistic	.350	.337	.359	.357	.357	.375	.365	.342	.352	.353	.344	.353	.359	.345
fantastical	.650	.663	.641	.643	.643	.625	.635	.658	.648	.647	.656	.648	.641	.655
rugged	.480	.484	.480	.471	.457	.460	.472	.465	.472	.474	.459	.478	.470	.477
delicate	.520	.516	.520	.529	.543	.540	.528	.535	.528	.526	.541	.522	.530	.523
sharp	.528	.518	.520	.523	.534	.514	.515	.499	.500	.518	.500	.512	.512	.513
round	.472	.482	.480	.477	.466	.485	.485	.501	.500	.482	.500	.488	.488	.486
simple	.517	.507	.520	.515	.515	.532	.517	.520	.511	.509	.519	.512	.524	.512
complex	.483	.493	.480	.485	.485	.468	.483	.480	.489	.491	.481	.488	.476	.488
solid	.358	.373	.364	.362	.362	.383	.369	.357	.363	.360	.354	.370	.374	.367
nonsolid	.642	.627	.636	.638	.638	.617	.631	.643	.637	.640	.646	.630	.626	.633
strong	.504	.506	.502	.501	.515	.503	.497	.487	.502	.504	.499	.507	.504	.501
weak	.496	.494	.498	.499	.485	.497	.503	.513	.498	.496	.501	.493	.496	.499
structured	.328	.340	.337	.343	.331	.356	.338	.343	.337	.334	.346	.341	.349	.341
disorganized	.672	.660	.663	.657	.669	.644	.662	.657	.663	.666	.654	.659	.651	.659
tense	.507	.508	.493	.506	.502	.502	.495	.497	.504	.510	.504	.515	.499	.497
relaxed	.493	.492	.507	.494	.498	.498	.505	.503	.496	.490	.496	.485	.501	.503

Table 19: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - Audio - Part 3).

Dimension	l (retroflex)	r	ɾ	j	ɕ	ʑ	ʔ	w	ɥ
abrupt continuous	.523 .477	.525 .475	.513 .487	.516 .484	.513 .487	.514 .486	.518 .482	.523 .477	.528 .472
active passive	.498 .502	.508 .492	.497 .503	.504 .496	.511 .489	.510 .490	.514 .486	.504 .496	.491 .509
beautiful ugly	.462 .538	.483 .517	.488 .512	.490 .510	.468 .532	.449 .551	.499 .501	.466 .534	.431 .569
big small	.496 .504	.519 .481	.479 .521	.495 .505	.498 .502	.492 .507	.505 .495	.500 .500	.501 .499
dangerous safe	.527 .472	.530 .470	.541 .459	.515 .485	.506 .494	.520 .480	.512 .488	.505 .495	.515 .485
exciting calming	.501 .499	.496 .504	.482 .518	.487 .513	.495 .505	.484 .515	.500 .500	.495 .505	– –
fast slow	.474 .526	.485 .515	.459 .541	.473 .527	.489 .511	.450 .550	.506 .494	.484 .516	.451 .549
good bad	.493 .507	.497 .503	.487 .513	.501 .498	.488 .512	.482 .518	.496 .504	.494 .506	.414 .586
happy sad	.484 .516	.483 .517	.475 .525	.487 .513	.497 .503	.515 .485	.491 .508	.500 .500	– –
hard soft	.502 .497	.499 .501	.490 .510	.484 .515	.493 .507	.483 .517	.485 .515	.488 .512	.494 .506
harsh mellow	.308 .692	.316 .684	.306 .694	.288 .712	.309 .691	.303 .697	.278 .722	.303 .697	.258 .742
heavy light	.515 .485	.517 .483	.500 .499	.508 .492	.529 .471	– –	.508 .492	.507 .493	.505 .495
inhibited free	.499 .500	.479 .521	.496 .504	.494 .506	.505 .495	.509 .490	.495 .505	.496 .504	.479 .521
interesting uninteresting	.501 .499	.513 .487	.463 .537	.504 .496	– –	– –	.504 .496	.507 .493	– –
masculine feminine	.515 .485	.508 .492	.476 .523	.486 .514	.492 .508	– –	.498 .502	.503 .497	– –
pleasant unpleasant	.488 .512	.499 .501	.498 .502	.500 .500	.490 .510	.464 .536	.507 .493	.494 .506	.506 .494
realistic fantastical	.347 .653	.349 .651	.321 .679	.337 .663	.346 .654	.333 .667	.369 .631	.354 .646	.342 .658
rugged delicate	.473 .527	.494 .506	.470 .530	.466 .534	.472 .528	.492 .508	.468 .532	.472 .527	.458 .542
sharp round	.515 .485	.523 .477	.501 .499	.513 .487	.546 .454	.520 .480	.525 .475	.513 .487	.521 .479
simple complex	.501 .499	.510 .490	.509 .491	.520 .480	.516 .484	.499 .501	.521 .479	.515 .485	.515 .485
solid nonsolid	.378 .622	.374 .626	.358 .642	.364 .636	.376 .624	.371 .629	.373 .627	.374 .626	.395 .604
strong weak	.501 .499	.509 .491	.497 .503	.500 .500	.513 .487	.477 .523	.506 .494	.497 .503	.489 .511
structured disorganized	.343 .657	.345 .655	.341 .659	.337 .663	.350 .650	.335 .665	.345 .655	.338 .662	.343 .657
tense relaxed	.509 .490	.502 .498	.505 .495	.492 .508	.529 .471	.512 .488	.500 .500	.494 .506	.511 .489

Table 20: Phoneme-Semantic Feature Attention Fraction Scores (Natural words - Audio - Part 4).

Dimension	i	ej	ɑ	ow	b	d	ɡ	p	t	k	v	ð	z	f
abrupt	.513	.527	.514	.513	.527	.533	.527	.506	.497	.511	.465	.460	.491	.521
continuous	.487	.473	.486	.486	.473	.467	.472	.494	.503	.489	.535	.540	.509	.479
active	.498	.498	.485	.487	.498	.509	.497	.495	.500	.506	.504	.496	.502	.498
passive	.501	.501	.514	.512	.501	.491	.502	.504	.500	.494	.496	.504	.497	.502
beautiful	.522	.534	.529	.511	.434	.410	.408	–	–	–	–	–	–	.543
ugly	.478	.465	.471	.489	.565	.590	.592	–	–	–	–	–	–	.456
big	.463	.439	.586	.567	.550	.555	.547	.521	.508	.505	.556	.527	.541	.491
small	.537	.561	.414	.433	.450	.445	.453	.479	.492	.494	.444	.473	.459	.509
dangerous	.513	.513	.510	.510	–	–	–	.522	.528	.531	.527	.530	.526	–
safe	.486	.487	.490	.489	–	–	–	.478	.471	.469	.472	.470	.474	–
exciting	.509	.524	.465	.480	.498	.499	.496	.521	.529	.531	.533	.482	.527	.512
calming	.491	.475	.534	.519	.501	.501	.503	.478	.471	.469	.467	.518	.473	.488
fast	.548	.551	.453	.438	.493	.484	.490	.555	.555	.547	.556	.537	.540	.494
slow	.451	.449	.547	.562	.507	.516	.510	.445	.445	.453	.444	.463	.460	.506
happy	.567	.577	.412	.425	.485	.476	.480	.495	.485	.487	.484	.500	.476	.492
sad	.432	.423	.588	.574	.514	.524	.520	.505	.514	.513	.516	.500	.524	.508
hard	.482	.484	.488	.467	.494	.488	.485	.534	.518	.519	.492	–	.525	.485
soft	.517	.516	.512	.532	.506	.512	.515	.466	.482	.480	.507	–	.475	.515
harsh	.262	.281	.267	.288	.283	.284	.258	.350	.348	.337	.350	.342	.340	.288
mellow	.738	.719	.733	.712	.717	.716	.742	.650	.652	.663	.650	.658	.660	.712
heavy	.452	.443	.543	.539	.560	.553	.541	.511	.506	.512	.556	.527	.537	.480
light	.548	.557	.457	.461	.440	.447	.459	.489	.494	.488	.444	.473	.463	.520
inhibited	.497	.474	.468	.467	–	–	–	.560	.564	.561	–	–	–	–
free	.503	.525	.532	.533	–	–	–	.440	.436	.439	–	–	–	–
masculine	.472	.472	.497	.477	.506	.487	.505	.504	.500	.493	.497	.481	.500	.478
feminine	.528	.528	.503	.523	.494	.513	.495	.496	.500	.506	.503	.519	.499	.521
realistic	.335	.328	.366	.366	.379	.400	.348	.340	.325	.290	.381	.378	.384	.362
fantastical	.665	.672	.634	.634	.621	.600	.652	.660	.675	.710	.619	.622	.616	.638
rugged	.464	.462	.478	.478	.505	.487	.499	.460	.439	.460	.416	.401	.420	.485
delicate	.536	.538	.522	.521	.494	.513	.500	.540	.561	.540	.584	.598	.580	.514
sharp	.537	.541	.486	.482	.498	.492	.493	.542	.543	.543	.554	.508	.554	.523
round	.462	.459	.513	.518	.501	.507	.506	.458	.456	.456	.445	.492	.445	.477
simple	.519	.520	.512	.526	.522	.519	.499	.519	.514	.504	.433	–	.434	.514
complex	.481	.480	.488	.473	.478	.481	.500	.481	.485	.495	.567	–	.565	.486
solid	.358	.360	.362	.364	.365	.357	.371	.422	.428	.466	.362	.359	.356	.354
nonsolid	.642	.640	.638	.636	.635	.643	.629	.578	.571	.534	.638	.641	.643	.646
strong	.495	.482	.486	.497	–	–	–	–	–	–	.556	.563	.571	–
weak	.505	.518	.514	.503	–	–	–	–	–	–	.444	.437	.429	–
tense	.505	.513	.500	.501	–	–	–	.566	.578	.556	–	–	–	–
relaxed	.494	.487	.500	.499	–	–	–	.434	.421	.444	–	–	–	–

Table 21: Phoneme-Semantic Feature Attention Fraction Scores (Constructed words - IPA text - Part 1).

Dimension	s	ʃ	m	n	l
abrupt	.516	.513	.515	.529	.515
continuous	.484	.487	.485	.471	.485
active	.494	.498	.481	.500	.489
passive	.505	.502	.519	.500	.511
beautiful	.556	.529	.551	.553	.537
ugly	.444	.471	.449	.447	.463
big	.478	.481	.463	.465	.462
small	.522	.519	.537	.535	.538
dangerous	–	–	.456	.455	.477
safe	–	–	.544	.545	.523
exciting	.505	.511	.480	.486	.481
calming	.495	.489	.519	.514	.519
fast	.478	.477	.439	.447	.454
slow	.522	.523	.560	.553	.545
happy	.479	.483	.488	.495	.483
sad	.521	.517	.511	.504	.517
hard	.481	.476	.488	.483	.483
soft	.518	.524	.512	.517	.517
harsh	.283	.279	.267	.275	.280
mellow	.717	.721	.733	.725	.720
heavy	.465	.460	.444	.438	.441
light	.534	.540	.556	.562	.559
inhibited	–	–	.444	.450	.439
free	–	–	.556	.550	.561
masculine	.488	.483	.468	.498	.500
feminine	.512	.517	.532	.501	.500
realistic	.379	.351	.366	.384	.377
fantastical	.621	.649	.634	.616	.623
rugged	.488	.476	.482	.494	.482
delicate	.511	.523	.518	.505	.518
sharp	.515	.519	.486	.473	.483
round	.485	.481	.514	.527	.517
simple	.522	.511	.515	.520	.524
complex	.477	.489	.485	.480	.475
solid	.334	.359	.359	.345	.367
nonsolid	.666	.641	.641	.655	.633
strong	–	–	.454	.470	.476
weak	–	–	.546	.530	.524
tense	–	–	.486	.475	.481
relaxed	–	–	.514	.524	.519

Table 22: Phoneme-Semantic Feature Attention Fraction Scores (Constructed words - IPA text - Part 2).

Dimension	i	ɑ	b	d	g	p	t	k	v	z	f	s	ʃ	m
abrupt	.522	.539	.531	.521	.529	.505	.513	.521	.512	.520	.526	.523	.529	.526
continuous	.478	.461	.469	.479	.471	.495	.486	.479	.488	.480	.474	.477	.471	.474
active	.507	.510	.516	.511	.509	.515	.509	.510	.513	.506	.506	.513	.508	.515
passive	.493	.490	.484	.489	.491	.485	.491	.490	.487	.494	.494	.487	.491	.485
beautiful	.511	.505	.477	.474	.477	–	–	–	–	–	.522	.523	.518	.517
ugly	.489	.495	.523	.526	.523	–	–	–	–	–	.478	.477	.482	.483
big	.499	.536	.521	.514	.517	.502	.501	.498	.516	.507	.503	.493	.499	.498
small	.501	.464	.479	.486	.483	.498	.499	.502	.484	.493	.497	.507	.501	.502
dangerous	.509	.517	–	–	–	.502	.507	.504	.511	.514	–	–	–	.552
safe	.491	.483	–	–	–	.498	.493	.495	.489	.486	–	–	–	.448
exciting	.510	.498	.504	.506	.509	.496	.520	.504	.511	.521	.515	.511	.515	.499
calming	.490	.502	.496	.494	.491	.504	.480	.496	.489	.479	.485	.489	.485	.501
fast	.525	.465	.492	.490	.491	.519	.536	.530	.519	.528	.495	.488	.481	.467
slow	.475	.535	.508	.510	.509	.481	.464	.470	.481	.472	.505	.512	.519	.533
happy	.508	.476	.506	.499	.498	.503	.503	.506	.505	.496	.508	.487	.497	.499
sad	.492	.524	.494	.500	.502	.497	.497	.494	.495	.504	.491	.513	.503	.501
hard	.490	.489	.489	.488	.484	.504	.510	.520	.520	.519	.467	.469	.473	.487
soft	.510	.511	.511	.512	.516	.496	.490	.480	.480	.481	.533	.531	.527	.513
harsh	.282	.297	.275	.278	.272	.319	.319	.314	.315	.319	.280	.267	.277	.265
mellow	.718	.703	.725	.722	.728	.681	.681	.686	.685	.681	.721	.733	.723	.735
heavy	.475	.522	.519	.519	.518	.497	.496	.503	.526	.519	.485	.480	.484	.463
light	.525	.478	.481	.481	.482	.503	.504	.497	.474	.481	.515	.520	.516	.537
inhibited	.475	.472	–	–	–	.473	.493	.494	–	–	–	–	–	.461
free	.525	.528	–	–	–	.527	.506	.506	–	–	–	–	–	.539
masculine	.487	.509	.502	.500	.504	.502	.498	.507	.500	.503	.492	.501	.492	.492
feminine	.513	.491	.498	.500	.496	.498	.502	.493	.500	.497	.508	.499	.508	.508
realistic	.366	.356	.366	.372	.360	.361	.368	.365	.363	.373	.354	.374	.367	.364
fantastical	.634	.644	.634	.628	.640	.639	.632	.635	.637	.627	.646	.626	.633	.636
rugged	.480	.498	.508	.491	.496	.476	.476	.481	.474	.481	.479	.504	.475	.490
delicate	.520	.502	.492	.509	.504	.524	.524	.519	.526	.519	.521	.496	.525	.510
sharp	.535	.505	.507	.503	.495	.533	.537	.533	.534	.541	.524	.524	.530	.498
round	.465	.495	.493	.497	.504	.467	.463	.467	.466	.458	.476	.476	.470	.501
simple	.517	.515	.515	.518	.512	.518	.516	.511	.471	.466	.516	.519	.515	.516
complex	.483	.485	.485	.482	.488	.482	.484	.489	.529	.534	.484	.481	.485	.484
solid	.356	.367	.350	.352	.351	.387	.392	.392	.359	.349	.359	.349	.345	.351
nonsolid	.644	.633	.650	.648	.649	.613	.608	.608	.641	.651	.641	.651	.655	.649
strong	.514	.500	–	–	–	–	–	–	.516	.521	–	–	–	.494
weak	.486	.500	–	–	–	–	–	–	.484	.479	–	–	–	.506
tense	.513	.510	–	–	–	.500	.504	.503	–	–	–	–	–	.511
relaxed	.487	.490	–	–	–	.499	.496	.497	–	–	–	–	–	.489

Table 23: Phoneme-Semantic Feature Attention Fraction Scores (Constructed words - Audio - Part 1).

Dimension	n	l
abrupt	.522	.524
continuous	.478	.476
active	.513	.518
passive	.487	.482
beautiful	.530	.514
ugly	.470	.486
big	.499	.501
small	.501	.499
dangerous	.558	.535
safe	.442	.465
exciting	.503	.499
calming	.497	.500
fast	.456	.462
slow	.544	.538
happy	.496	.495
sad	.504	.505
hard	.482	.479
soft	.518	.521
harsh	.259	.272
mellow	.741	.728
heavy	.463	.462
light	.537	.538
inhibited	.444	.456
free	.556	.544
masculine	.495	.496
feminine	.505	.504
realistic	.367	.356
fantastical	.633	.644
rugged	.498	.487
delicate	.502	.513
sharp	.489	.497
round	.511	.503
simple	.519	.521
complex	.481	.479
solid	.349	.354
nonsolid	.651	.646
strong	.491	.492
weak	.509	.508
tense	.511	.493
relaxed	.489	.507

Table 24: Phoneme-Semantic Feature Attention Fraction Scores (Constructed words - Audio - Part 2).

Layer	Natural		Constructed	
	IPA	**Audio**	**IPA**	**Audio**
0	0.5022	0.4974	0.5151	0.4950
1	0.5047	0.5000	0.5048	0.4901
2	0.5060	0.5023	0.5109	0.5000
3	0.5054	0.5002	0.5147	0.4982
4	0.5020	0.5036	0.5100	0.5036
5	0.5090	0.5018	0.5228	0.5052
6	0.5026	0.4985	0.5022	0.4845
7	0.5050	0.5007	0.5394	0.5152
8	0.5087	0.5049	0.5395	0.5185
9	0.5121	0.5076	0.5110	0.5045
10	0.5112	0.5030	0.5253	0.5083
11	0.5096	0.5022	0.5311	0.5088
12	0.5064	0.4957	0.5409	0.5036
13	0.5061	0.4964	0.5270	0.5049
14	0.5065	0.4962	0.5269	0.5111
15	0.5077	0.5046	0.5282	0.5048
16	0.5140	0.5022	0.5249	0.4974
17	0.5075	0.5011	0.5180	0.5023
18	0.5065	0.5058	0.5291	0.5144
19	0.5092	0.5002	0.5062	0.5019
20	0.5055	0.5056	0.5165	0.5104
21	0.5091	0.5016	0.5186	0.5092
22	0.5119	0.5002	0.5285	0.5095
23	0.5082	0.4985	0.5229	0.5148
24	0.5049	0.4991	0.5427	0.5116
25	0.5077	0.4998	0.5337	0.5101
26	0.5104	0.4983	0.5324	0.5076
27	0.5056	0.4987	0.5101	0.5043
Average	**0.5073**	**0.5009**	**0.5226**	**0.5053**

Table 25: Layer-wise attention fraction scores for IPA and audio input types across natural and constructed word groups in Qwen2.5-Omni-7B.

You are a professional linguistic annotator.
Please read a `{language}` mimetic word and its meaning, and decide which semantic feature best describes the word's meaning.

[WORD]
`{word}`

[MEANING]
`{meaning}`

[SEMANTIC DIMENSION]
`{feature1}` vs. `{feature2}`

[OPTIONS]
1: `{feature1}`
2: `{feature2}`
3: Neither
Answer with the number only. (1-3)

Table 26: Prompt for automatic semantic dimension annotation.

Given a [WORD], which semantic feature best describes the word based on auditory impression?

[WORD]
`{word}`

[SEMANTIC DIMENSION]
`{feature1}` vs. `{feature2}`

[OPTIONS]
1: `{feature1}`
2: `{feature2}`
Answer with the number only. (1-2)

(a) Prompt for original text words.

Given an IPA [WORD], which semantic feature best describes the word based on auditory impression?

[WORD]
`{word}`

[SEMANTIC DIMENSION]
`{feature1}` vs. `{feature2}`

[OPTIONS]
1: `{feature1}`
2: `{feature2}`
Answer with the number only. (1-2)

(b) Prompt for IPA text words.

Given a spoken [WORD], which semantic feature best describes the word based on auditory impression?

[WORD]
`{audio}`

[SEMANTIC DIMENSION]
`{feature1}` vs. `{feature2}`

[OPTIONS]
1: `{feature1}`
2: `{feature2}`
Answer with the number only. (1-2)

(c) Prompt for spoken audio words.

Table 27: Prompts for the semantic dimension prediction experiment by input type.