Universiteit Leiden

Magnitude symbolism without sound: investigating phonetic size-symbolic associations in Large Language Models

Ewelina Kowalczyk (s3981282)

Supervisors:
Tessa Verhoef & Kiana Shahrasbi

Bachelor Thesis– Data Science & Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

June 30, 2026

Abstract

Magnitude sound symbolism is a well-studied, cross-linguistic human cognitive bias. It is an association of certain sounds with certain sizes, for example, an association of the vowel /ɪ/ with smallness. While it is well-studied in humans, emerging from acoustic, articulatory, and evolutionary reasons, it is recently also becoming an area of growing academic interest within computational models and artificial intelligence. Although sound symbolism is a multimodal phenomenon, this paper investigates whether size sound-symbolic associations can persist within textual data, without sound. The goal of this research is to understand the extent to which Large Language Models (LLMs) encode human-like magnitude sound symbolism. The experiment is based on a large-scale cross-linguistic human study. It uses 40 disyllabic nonce words as stimuli for size judgments which are evaluated across three phonetic features (vowel height, vowel backness, and obstruent voicing). The task is performed by three LLMs (gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B). The words are presented in three different input formats (word, IPA, and spaced IPA) and judged on various scales (1-4, 1-2, and 1-8). Token log-probabilities are extracted from models' answers to obtain a more detailed insight into their decisions and capture the small nuances of sound-symbolic associations. The results are analyzed using linear (mixed effects) models. The findings show that all three LLMs display some sound-symbolic associations, however, each model to a different degree. Gpt-4o-mini most accurately reflects human magnitude sound-symbolic biases, almost perfectly replicating all human phonetic mappings. DeepSeek-V3.1 also reflects human biases accurately, however, only partially. In contrast, the only significant effects exhibited by Qwen3.5-9B involve a reversal of the human vowel height trend. Moreover, performance is highly sensitive to scale and input format types. Most significant findings are found on a scale of 1 to 4 in the spaced IPA format where despite the small visible vowel disparity, model judgments become precise with narrow 95% confidence intervals. These findings suggest that human-like magnitude sound-symbolic associations can be successfully extracted from textual data, given controlled and well-curated conditions. These results contribute to the debate about the capabilities and limitations of computational models in reflecting phenomena grounded in human cognition.

Contents

1 Introduction

1.1 Introduction to sound symbolism

The arbitrariness of languages is the leading assumption in human cognition (De Saussure et al., 1916; Hockett and Hockett, 1960). It claims that the relationship between words and their meanings is random and based on social agreement and convention instead of inherent logic. There is nothing about the word "tree" that makes it more appropriate to describe a tree than any other word. This argument is further strengthened by the fact that the same thing is denoted by different words in different languages, despite representing the exact same object.

However, there are multiple exceptions to the argument of the arbitrariness of languages. For example, iconicity is a concept that refers to a non-arbitrary link between a form of a written word, a spoken sound, or a gesture and its meaning (Dingemanse et al., 2015). A common example of iconicity in many languages is onomatopoeia; words such as 'bang' in English describing a loud noise, or 'boe' in Dutch describing a noise made by a cow, or 'apsik' in Polish describing sneezing. All of these words sound like the things (usually animals and objects) they are describing, helping to convey meaning and making them easier to understand. Onomatopoeia is an example of direct iconicity, linking a word to the exact meaning. However, there is also sound symbolism, an example of indirect iconicity, that allows to form *associations* between spoken sounds and meanings (Sidhu and Pexman, 2018).

Sound symbolism is a phenomenon in which speech sounds of words (phonemes) are associated with certain meanings. The associations mainly refer to more general, descriptive, perceptual features of the objects, such as size or roundness. Some of the more well-known types of sound symbolism include shape symbolism and magnitude symbolism. Shape symbolism, known as the bouba-kiki effect (or maluma-takete, as introduced by Köhler (1929)) is an association between phonetic features of words and the roundness/sharpness of the object being described. People tend to associate nonce words such as *bouba* or *maluma* with rounder and smoother shapes, whereas words such as *kiki* or *takete* are often matched with sharper shapes (Ramachandran and Hubbard, 2001), as shown in **Figure 1**.

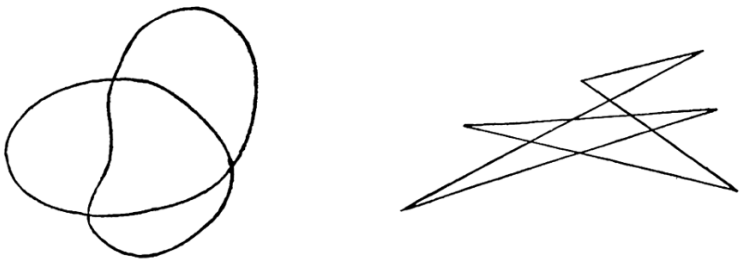


Figure 1: The first shape is often associated with words such as *bouba* while the second shape is more likely to be associated with words such as *kiki*. Images from Köhler (1929); Köhler (1947).

On the other hand, magnitude symbolism, also called size symbolism or mil-mal effect, is the correlation between phonetic features of words and the size associated with what these words describe, which was originally introduced by Sapir (1929). For example, words containing front-high

vowels such as *i* are more likely to be judged as smaller, while words containing back vowels such as *a* or *o* are usually associated with bigger sizes. Additionally, there are other examples of sound symbolism phenomena such as associations between vowel height and frontness and perceptual features such as color and lightness studied for example by Cuskley et al. (2019). Despite focusing on one language, Dutch, the study found strong evidence that these associations are shared across the majority of individuals (around 70%), supporting the claim that some relationships between sounds and perceptual features are non-arbitrary.

Larger studies show that sound symbolism seems to be rooted in languages all across the world. Blasi et al. (2016) analyzed 100 words in 4,298 languages and found, among other non-arbitrary mappings, a clear relation between smallness and the vowel *i*. This goes to show that sound symbolism is not language-specific, but may rather be a systematic phenomenon in human cognition. While arbitrariness in language is an important feature because it allows for an unlimited number of words to be created, not constrained by the need of having a direct or indirect association to its meaning, sound symbolism, and even more generally iconicity, makes the language more graphic and direct (Lockwood and Dingemanse, 2015). Moreover, sound symbolism has been shown to help with learning the vocabulary of a language, both in children and adults (Imai et al., 2008).

As shown, language is not completely arbitrary, but rather both arbitrariness and iconicity play an important role. Research into the non-arbitrary parts of language, such as sound symbolism, reveals interesting truths about human cognition. Sound symbolism is supported by multiple empirical studies that demonstrate consistent cross-linguistic associations between phonetic features and stimuli across various modalities.

1.2 Magnitude sound symbolism in text-only modality: from humans to Large Language Models

While the exact reasons why humans display sound-symbolic associations are not known, it is clear that this phenomenon is cross-modal and phonetically grounded. Magnitude sound symbolism makes a connection between auditory and visual modalities, which are deeply embedded in human experience. However, this raises the question of whether these biases can be displayed without the auditory grounding and experience. This paper aims to provide insight into how well size sound-symbolic associations survive a transition into a text-only medium. Testing this phenomenon on Large Language Models (LLMs) allows for exploration of the strength with which these biases are transferred into text. Since LLMs are purely dependent on statistical patterns within text when processing language, their ability to display size sound-symbolic mappings could be used as an argument that auditory experience is not necessary to adopt these human-like cognitive biases. The findings of this paper will contribute to the ongoing debate about the extent to which computational models can exhibit sound-symbolic effects as well as provide further arguments regarding the origin of this phenomenon.

1.3 Research Question

Thereby the research question is:
To what extent do LLMs encode human-like sound-size symbolism?
Several subquestions will be used to facilitate the process of answering the main RQ, including:
RQ1: In which of the researched LLMs does the magnitude sound-symbolism show up?

RQ2: Do LLMs show similar phonetic associations between words and their imagined size as humans?
RQ3: What does this suggest about the origin of the size sound-symbolism effect?

1.4 Overview of methodological approach

To answer the RQs, this paper will evaluate three different LLMs (gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B) on a task in which words need to be judged based on their perceived largeness.

The methodology of the experiment will be based on a cross-linguistic study on 103 human participants by Shinohara and Kawahara (2010). The study investigated the magnitude sound symbolism on speakers of Chinese, English, Japanese, and Korean. The participants were given instructions: *"Imagine an exotic language that you don't know. The language has a rich lexical inventory of adjectives that express a variety of "largeness" or "smallness". Now, a speaker of this language looks inside a box and finds a jewel. She verbally expresses how large or small it looks using one of these adjectives. Your task is to read each of the following words and guess its meaning — i.e., how large or small it is."* and were asked to judge the size of 40 disyllabic nonce words on a scale of 1 to 4. The study noted that in general, participants rated *i* as smaller than other vowels, while *a* and *o* were typically rated as larger. Moreover, the study found significant results that height, backness, and voicing all affect the size judgments. Among all languages, the most common trends showed that back vowels are rated as larger than front vowels, voiced obstruents are rated as larger than voiceless obstruents, and, for the most part, the lower the height of the vowel, the larger the size rating. The study tried to find articulatory and acoustic explanations of the phenomenon and noted that both are possible and could be the reason why certain phonetic features are so consistently associated with size judgments.

The chosen LLMs will be prompted just like humans in the study by Shinohara and Kawahara (2010) to see if they exhibit size sound-symbolic biases. The models will be asked to judge the size of 40 nonce words on various scales and presented in different formats to see how these changing conditions impact the robustness of size judgments. Token log-probabilities will be extracted from models' answers to get a more detailed view into models' decisions and to capture the nuance of sound symbolism effects.

The experiment will allow comparison of LLMs judgments to those of a large group of people across different languages. In this way, the phonetic features that affect size judgments will be compared in detail. Based on similarities and differences found in judgments, it will be inferred whether human-like size sound-symbolic associations can emerge without any sound and whether they can be expressed through biased textual data alone.

2 Background and Related Work

2.1 Sound symbolism in human studies

Sound symbolism has been extensively studied in humans, with shape sound symbolism receiving a lot of academic attention. The bouba-kiki effect has been analyzed and observed by Ćwiek et al. (2022) in 25 different languages representing 10 writing systems, providing strong evidence that this effect is valid cross-culturally. Moreover, a study by Maurer et al. (2006) investigated the

bouba-kiki effect in children and compared it to adults, reporting that no significant differences were found between the decisions of the two age groups, noting that the effect is independent of age. Both of these findings suggest that shape symbolism is naturally embedded in human cognition.

Interestingly, multiple studies have demonstrated that sound-symbolic associations are stored implicitly in the brain and processed automatically. Preziosi and Coane (2017) investigated how congruent and incongruent pairings of sound and perceived size impact memory and recall. The findings showed that congruent pairs (previously matched by human participants) were easier to remember, suggesting implicit encoding of size symbolic associations. In addition, Parise and Spence (2012) investigated response times in matching takete-maluma and bouba-kiki auditory stimuli with congruent and incongruent visual stimuli. The study found that the congruent pairs elicited faster responses, suggesting that these cross-modal associations are processed non-consciously.

2.2 Magnitude sound symbolism in human studies

Besides the large scale study by Blasi et al. (2016) that found evidence of magnitude symbolism across multiple languages, other studies have also researched the size symbolism phenomenon. All of them found that humans tend to associate smaller objects with words that contain vowels such as *i* and larger objects with words that contain vowels such as *a* (Berlin, 2006; Newman, 1933; Sapir, 1929).

Moreover, interestingly, some studies reveal not only that some words are associated with big objects and others with small objects, but that these size judgments can even be graded on a scale. Thompson and Estes (2011) conducted an experiment in which participants were asked to pick a name that describes an object of varying size from a list of made-up words. The researchers discovered a linear relationship between the size of the object and the number of large-sounding phonemes present in the word chosen by the participants. This goes on to show that the magnitude of sound-symbolic judgments are not only systematic but also scalable.

Shinohara and Kawahara (2010) tried to give possible acoustic and articulatory explanations as to why these mappings occur. Within articulatory explanations they pointed to how phonetic features associated with larger sizes are connected to enlarged parts of human speech system. For example, the lower the vowel, the wider the aperture when the vowel is pronounced. Moreover, when spoken, both back vowels and voiced obstruents enlarge sub-oral cavity. On the other hand, acoustic explanations rely on the frequency code hypothesis which ties the size of a resonator to the resulting frequency. Lower frequencies are produced by larger resonators and vice versa. This inverse relationship between judged size and frequency (specifically second resonance frequency F2) may be an explanation as to why certain vowels are judged as smaller than others. Moreover, pairing vowels with voiced obstrunets decreases their fundamental frequency (F0), which again may cause larger size judgments. While both articulatory and acoustic reasons could explain the emergence of magnitude sound symbolism, the exact origins of this phenomenon are not fully known.

Although it is clear that sound symbolism is a legitimate phenomenon among humans, knowing more about its origins could provide insight into how humans understand and process language. While interest in investigating sound symbolism, its role, and the reasons for its emergence is growing (Nielsen and Dingemanse, 2021; Fischer et al., 2026), the exact cognitive explanations of this phenomenon are not known, making it an interesting field to research. Finding a clear explanation of sound symbolism could further specify how much of this phenomenon is embedded

in human cognition, dependent on lived experiences, or rooted in language structure.

2.3 Possible cognitive explanations of sound symbolism

Although there is no definitive justification for sound symbolism, the review by Sidhu and Pexman (2018) outlines five main mechanisms that could serve as explanations for the existence of this phenomenon in humans. All of them are conceptualized as "arising from associations between specific phonetic features and particular perceptual and/or semantic features."(Sidhu and Pexman, 2018, p. 1624) This already suggests that sound symbolism is not only embedded in human cognition, but also extended through multi-modal experiences of the world. According to the review, the five mechanisms of sound symbolism include statistical co-occurrence, shared properties, neural factors, species-general associations, and language patterns.

Statistical co-occurrence represents a high probability of two associated stimuli appearing together in the environment. It is said to explain, for example, the associations of high-pitched sounds with small sizes and low-pitched sounds with large sizes (Gallace and Spence, 2006). This is a common occurrence in the world, as larger objects/bodies tend to produce louder, and thus lower, sounds (Spence, 2011).

Another potential mechanism, shared properties, presumes that some associations may be formed because sound and the other corresponding stimulus have common properties, despite being derived from different modalities. The properties may either be perceptual, such as the experience of size, or more conceptual, derived from connotative meanings of stimuli. This explanation could justify the sound symbolic associations between high-front vowels and coldness, given that coldness and smallness share a similar connotation, and smallness is associated with high-front vowels (French, 1977).

The third mechanism, neural factors, is based specifically on the way information is processed in the brain. For instance, one theory (Walsh, 2003) suggests that the magnitude of the stimulus, regardless of the modality, might be encoded similarly, leading to associations between, for example, high volume and brightness (both represent high values on a magnitude scale) (Marks, 1987; Spence, 2011).

Species-general associations, the fourth suggested mechanism explaining sound symbolism, claims that cross-modal associations are shared across species and have become embedded into human nature as a result of evolution. An example of such a view is frequency code theory introduced by Ohala (1995) suggesting that high-frequency sounds are associated with small objects, while low-frequency sounds are associated with big objects. This has its applications in nature, for example with animals using low-frequency noises to appear threatening. While this view is very similar to the first mechanism, statistical co-occurrence, it is argued that these associations are evolutionarily ingrained rather than learned through experience.

The last group of mechanisms, language patterns, suggests that sound-symbolic associations arise from phonological and semantic features that tend to coexist in language structure. For example, if some phonemes are commonly used in words that convey similar meanings, these phonemes may become associated with these general semantic features. However, such a theory can only explain certain examples of sound symbolism, making additional explanations necessary to grasp the origins of the entire phenomenon.

As described above, the first two mechanisms seem to be explaining sound symbolism as emergent from experience and interacting with the world. In contrast, the third and fourth mechanisms

suggest that sound symbolism is embedded in human nature and cognition. Conversely, the last mechanism claims that this phenomenon is rooted in language. It seems like none of the mechanisms can fully explain where sound symbolism emerges from, but rather each one of them matches a subset of sound-symbolic association examples to some intuitive explanation. This open interpretation raises a question about prerequisites of forming sound-symbolic associations and whether sound is a necessary component after all.

2.4 Sound symbolism without sound

There is a chance that sound-symbolic effects may arise without sound. After all, language has been shaped by generations, with spoken and written forms closely intertwined. Therefore, it is possible for the associations between certain sounds and stimuli of other modalities to be adopted purely based on written language.

This hypothesis could potentially be tested on humans simply by conducting sound-symbolic association experiments without auditory input, only providing the stimulus in written form. However, this approach has serious limitations. People could still read the words 'in their head', leading to an internally generated experience of sound (Leinenger, 2014). While this could still be useful in determining whether the sound needs to be actually heard or if it is enough for it to just be imagined, it does not provide insight into whether these associations can emerge when auditory processing is completely lacking.

Investigating sound symbolism with, for example, Large Language Models (LLMs) and Vision Language Models (VLMs) can provide useful insight into this problem. If these computational models turn out to show sound-symbolic associations, it would potentially mean that they have learned them from statistical structures embedded in the written language and data in the training corpus.

2.5 Sound symbolism in computational models

Perhaps one of the most recent and extensive research on sound symbolism with computational models is that by Jeong et al. (2026). The study used Multimodal Large Language Models (MLLMs) and investigated sound-symbolic associations across 19 different dimensions, including shape, size, speed, strength, weight, and many more. The choice of multimodal models allowed for the examination of the importance of sound, providing useful analysis into the nature of the phenomenon. The models were prompted with words in written form to provide a baseline, allowing the impact of the auditory modality to be assessed when the models were later prompted with sound. Importantly, the prompts included both natural and newly constructed words to test real generalizations, beyond what the models could have learned and memorized from the training data. The findings suggest that the presence of sound strengthens associations between phonetic features and meanings. Sound turned out to be especially influential for dimensions such as size and speed, while it played a smaller role in dimensions such as shape and valence. Although this study reveals interesting truths about sound symbolism, it has a limitation of comparing MLLMs' results to votes of only 10 human participants. This highlights the need for experiments with computational models that compare sound-symbolic associations, especially these that receive less attention, with large-scale, cross-linguistic human studies.

2.5.1 Bouba-kiki in computational models

Bouba-kiki effect is the sound symbolism example most commonly investigated in computational models, especially multimodal Vision Language Models (VLMs). One study of the bouba-kiki effect in VLMs found strong evidence that this phenomenon persists in models such as CLIP and Stable Diffusion (Alper and Averbuch-Elor, 2023). However, more recent studies on this topic revealed that no significant evidence was found and suggested that results may vary depending on the characteristics of the models used (Verhoef et al., 2024; Kouwenhoven et al., 2025). This indicates that while sometimes these large AI models can show signs of shape sound-symbolic associations, this is not consistent across models and may depend, for example, on architecture. Thus, it may suggest that only having been trained on textual and visual information in the corpus, VLMs cannot always reliably learn shape sound-symbolic correlations.

2.5.2 Magnitude sound symbolism in computational models

While magnitude symbolism is relatively widely researched in humans, there seems to be little research on it with AI models, especially one that is based on large-scale experiments with humans. Loakman et al. (2024) investigated this effect, together with shape symbolism and iconicity ratings, using state-of-the-art VLMs and LLMs. The models were prompted using both standard and informed prompts, where in the informed prompt condition the models were notified that the task relates to sound symbolism. The study found evidence that the investigated models can indeed demonstrate similar sound-symbolic associations to humans, with higher performance observed when informed prompts were used. Nonetheless, the study also noted that in some cases the models showed clear disagreements with human judgments. This could suggest that some knowledge within the model training data may overwrite the associations between sounds and meanings that humans tend to make. Additionally, the study only compared the model results with the majority vote of 10 human participants (all with native English speaking proficiency), raising concerns about the generalizability of the findings. This again emphasizes the need for further research.

2.6 Research gap

Magnitude symbolism has been extensively researched in humans, drawing evident conclusions that people tend to associate some sounds with certain size judgments. However, most studies on sound symbolism with computational models involve investigating the bouba-kiki effect and use multimodal language models. Additionally, they often compare models' results to a small group of human participants. Therefore, the occurrence of the magnitude symbolism phenomenon based on purely textual information is understudied. It is unclear to what degree size-symbolic associations can be formed or adopted in the absence of auditory input. Therefore, investigating magnitude symbolism in LLMs could help gain more insight into the origins of this phenomenon and its embedding in human cognition. It could also clarify how LLMs encode information and whether human phonetic biases are reflected in the textual corpus on which they are trained. Consequently, this research examines whether LLMs encode human-like magnitude sound-symbolic associations.

3 Methodology

The source code, experimental and analysis scripts, as well as the stimuli and resulting data are publicly available on GitHub.[1]

3.1 Prompting

The methodology of the experiment was largely based on the methodology of the study by Shinohara and Kawahara (2010). Instead of people of different nationalities, the experiment presented in this thesis used three different models; gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B, which would simulate the decisions of a group of people. This was done by prompting the models with 40 nonce words shown in Figure 3. The prompt used was slightly modified to provide a clearer explanation to LLMs. The prompt was: *"Imagine an exotic language that you don't know. The language has a rich lexical inventory of adjectives that express a variety of largeness' or 'smallness'. Now, a speaker of this language looks inside a box and finds a jewel. She verbally expresses how large or small it looks using the adjective pronounced stimulus. Your task is to guess the meaning of stimulus — i.e., how large or small it is. Make the size judgment on a scale of 1 to 4 (1=very small, 2=relatively small, 3=relatively large, 4=very large). Reply with ONLY the digit (1, 2, 3, or 4)."* The prompt was later also modified to ask for judgments on a scale of 1 to 2. The modified part stated: *" Make the size judgment by choosing 1 or 2 (1=small, 2=large). Reply with ONLY the digit (1 or 2)."* Moreover, the experiment was also once run on a judgment scale of 1 to 8, with the modification being: *"Make the size judgment on a scale of 1 to 8 (1=extremely small, 2=very small, 3=small, 4=quite small, 5=quite large, 6=large, 7=very large, 8=extremely large). Reply with ONLY the digit (1, 2, 3, 4, 5, 6, 7, or 8)."* While 1-4 scale was the original scale based on the study by Shinohara and Kawahara (2010), the 1-2 scale was introduced to decrease the bias of models towards middle values. Forcing the models to pick between 1 (small) and 2 (large) aims at achieving a more polarized size judgment distribution, showing clearer sound-symbolic association nuances. Lastly, the 1-8 scale was introduced to see if extending the 1-4 scale would model a clearer, and less biased against extreme ratings, distribution within the middle values of the scale.

When prompting the models, the temperature was set to zero to gather the answers that are as deterministic as possible. The models were prompted to respond with one token only and to return top five token-level logarithmic probabilities (logprobs) of their answers. Since both DeepSeek-V3.1 and Qwen3.5-9B are reasoning models, meaning they return "*thinking*" tokens before the actual answer, the reasoning was disabled to force them to give a numerical size judgment answer as the first token. Based on returned logprobs, the probabilities for each size judgment were gathered and normalized. Later, weighted average scores for each stimulus were calculated. The reason for using logprobs was to gain a deeper insight into model's internal certainty and to better capture subtle nuances of sound symbolic associations.

3.2 Models

A few chosen LLMs are used in the experiment. This ensures that the existence of the effect can be assessed across models with different architectures and properties. The chosen models include gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B.

[1] See the repository at https://github.com/ewelinaenia/Bachelor-Thesis-2026

Although the exact architectural details of gpt-4o-mini are not known, the model is a scaled-down version of gpt-4o so it is natively multimodal. It was chosen for this experiment to serve as the industry State-of-the-Art standard. Gpt-4o-mini offers strong performance while being extremely affordable and having improved latency. It scores only slightly below gpt-4o on most evaluation benchmarks, achieving accuracy of 82.0% on MMLU, compared to gpt-4o's 88.7%.(OpenAI, 2024)

DeepSeek-V3.1 is a Mixture-of-Experts (MoE) model using Multi-head Latent Attention (MLA). It has a total of 671B parameters, but only activates 37B for each token. It was chosen to see if a MoE model can still make consistent size judgments comparable to human decisions. It is also a text-only model, so it may reveal if multimodal training is actually necessary to exhibit magnitude sound symbolism. It is also interesting to see how a model that manages math reasoning tasks exceptionally well performs on language tasks that heavily rely on human subconscious biases rather than strict logic.(Liu et al., 2024)

Lastly, Qwen3.5-9B is a much smaller model with only 9B parameters; however, due to its hybrid architecture (combining Gated Delta Networks with sparse Mixture-of-Experts) and extremely wide linguistic coverage, it provides insight into how a small-scale but highly-optimized model can handle human linguistic biases. It is also natively multimodal and supports 201 languages and dialects, making it a strong pick for a study where model's judgments are meant to be compared to human judgments across different languages (mostly non-western).(hug)

The chosen models represent different architectures, sizes, and specializations. They provide insight into what is needed in a model for human size sound-symbolic biases to emerge, helping to make inferences about the origins of this phenomenon.

3.3 Stimuli

The stimuli used were the 40 nonce words, written in three different formats: *word* (e.g. *ibib*), International Phonetic Alphabet or *IPA* (e.g. *ibib*\), and *spaced IPA* (e.g. *i b i b*\), making a total of 120 unique stimuli. In their study on sound symbolism, Jeong et al. (2026) included spaced IPA text tokens as input type to observe how much of the associations could be made because of trained token memorization. In this experiment, including *spaced IPA* word format is meant to force the model to pay special attention to each token (since now the letters are separated), instead of looking at the word as a whole. *IPA* format, on the other hand, is included as a control group and the binding format between *word* and *spaced IPA*. By including IPA formats it will be possible to make inferences about origins of sound symbolism in LLMs. Finding out the formats in which sound-symbolic associations persist could confirm or deny an assumption that these associations are learned mainly directly from the words appearing in the corpus.

The classification of vowels into voicing, height, and backness was also done based on the study by Shinohara and Kawahara (2010). **Figure 3** shows the classification of obstruents, while **Figure 2** shows the classification of vowels into high, mid, low, as well as front and back.

3.4 Statistical analysis

The results were statistically assessed using linear mixed effects models, as in the original study by Shinohara and Kawahara (2010). The analysis was run separately for every input format type across different models (and for different scales of size judgments). Voicing, height, and backness were treated as fixed effects, and model name was treated as random effect (**Equation 1**). Additionally,

	Front	Back
High	*i*	*u*
Mid	*e*	*o*
Low		*a*

Figure 2: Categorization of vowels in terms of height and backness.

linear model analyses were run separately for every model across different format types (**Equation 2**) to further analyze model-specific trends and find out which models show the strongest size sound-symbolic associations similar to humans. Lastly, linear model analyses were run for every model for each format type separately (**Equation 3**) to better understand how input format types affect model judgments. On top of all that, post-hoc analyses were run to compare three levels of vowel height. P-values were automatically adjusted for multiple comparisons using Tukey's Honestly Significant Difference (HSD) procedure through the `emmeans` package in R.

$$avg\_rating \sim voicing + height + backness + (1|model\_name) \tag{1}$$

$$avg\_rating \sim voicing + height + backness + format \tag{2}$$

$$avg\_rating \sim voicing + height + backness \tag{3}$$

Voiced	b	d	g	z
i	ibib	idid	igig	iziz
u	ubub	udud	ugug	uzuz
e	ebeb	eded	egeg	ezez
o	obob	odod	ogog	ozoz
a	abab	adad	agag	azaz
Voiceless	p	t	k	s
i	ipip	itit	ikik	isis
u	upup	utut	ukuk	usus
e	epep	etet	ekek	eses
o	opop	otot	okok	osos
a	apap	atat	akak	asas

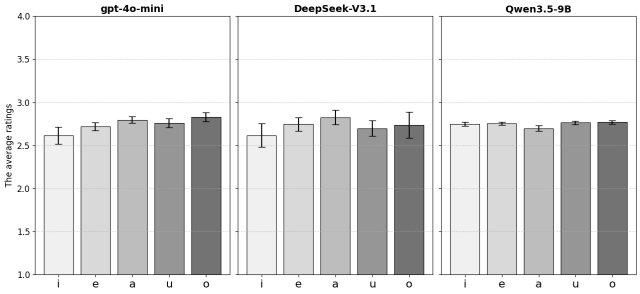
Figure 3: 40 nonce words from the study by Shinohara and Kawahara (2010) used to prompt LLMs for size judgments. The figure also shows the categorization of obstruents in terms of voicing.

4 Results

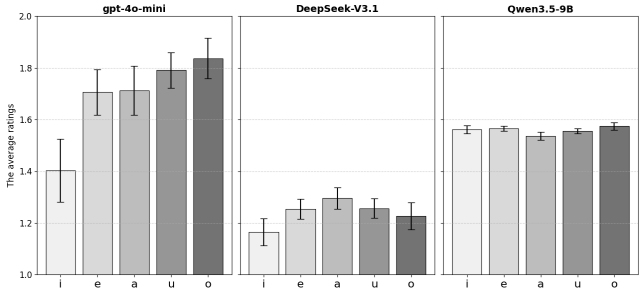
4.1 General trends

Figure 4 shows average ratings across all input format types for the five vowels for three types of models for (*a*)1-4, (*b*)1-2, and (*c*)1-8 size judgments, as well as (*d*)human judgments on a scale of 1-4 for comparison. The error bars represent 95% confidence intervals (here and in all the bar charts in this paper).

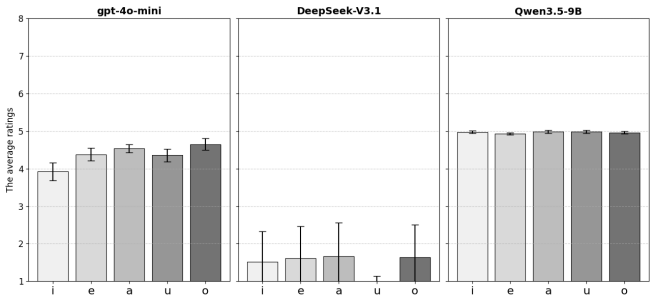
The majority of humans across the four languages in the study by Shinohara and Kawahara (2010) had a strong preference to judge words that contain *i* as the smallest, while words that contain *a* or *o* were frequently judged as the biggest. The differences between vowels were extreme enough to be significantly visible on the graph (4d). Meanwhile, in the same experiment conducted on LLMs on a scale of 1 to 4 the average ratings between vowels were much smaller and subtle (4a). While gpt-4o-mini and DeepSeek-V3.1 still showed slight preferences for juging *i* as smaller and *a* or *o* as larger, Qwen3.5-9B showed little to no variation between vowels, with only *a*, surprisingly, being judged as slightly smaller.



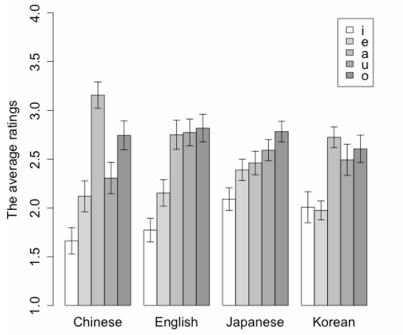
(a) Model size judgments on a scale of 1 to 4.



(b) Model size judgments on a scale of 1 to 2.



(c) Model size judgments on a scale of 1 to 8.



(d) Human size judgments on a scale of 1 to 4. Figure reproduced from Shinohara and Kawahara (2010).

Figure 4: Average ratings across words in all input format types containing one of the five vowels.

When prompted with a size judgment scale of 1 to 2, the average ratings became more polarized and, in some cases, more human-like (4b). Gpt-4o-mini showed a clear preference to classify words containing *i* as smaller. The rest of the vowels were classified as increasingly bigger, similarly to the trend of average vowel ratings in English and Japanese speakers, with the highest judged vowel being *o*. On the other hand, DeepSeek-V3.1 showed vowel size judgments similar to those

of Chinese and Korean speakers, with words containing vowel *i* remaining as the smallest, while words containing *a* were judged as the largest. On the contrary, Qwen3.5-9B judged vowel sizes completely differently, rating words containing *a* as the smallest. Again, the variation between vowels remained low for Qwen3.5-9B.

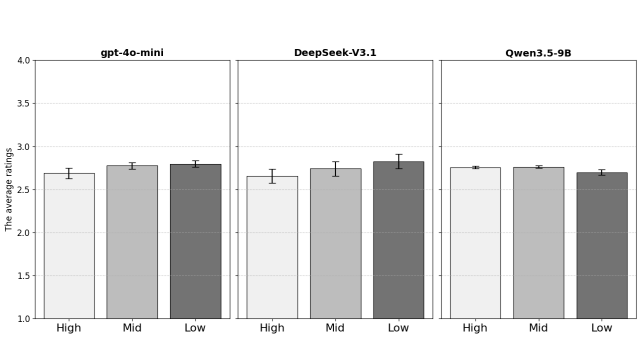
Lastly, when asked for size judgments on a scale of 1 to 8, all LLMs responded differently (4c). Gpt-4o-mini reverted to similar size ratings as for the scale of 1 to 4. The polarization obtained through the use of scale of 1 to 2 dissipated, presumably due to the model avoiding extremes of the scale (Rupprecht et al., 2025), aiming more for middle values. Conversely, DeepSeek-V3.1, for some reason, started to ignore instructions, often returning words or characters as initial tokens of the response. This led to multiple answers being deprived of any numbers, causing the final ratings of many words to be zero. This completely disrupted the results, making them useless in revealing the existence of sound-symbolic associations. Lastly, Qwen3.5-9B yet again judged the vowels extremely similarly, this time with *a* and *o* appearing to be rated as slightly smaller. Due to the results of judgments on a scale of 1 to 8 being very similar to those of judgments on a scale of 1 to 4 and partly compromised by the DeepSeek-V3.1 results, they were excluded from further statistical analysis.

It is worth noting that considering gpt-4o-mini and DeepSeek-V3.1 responses, there are some prominent agreements with human judgments. They include clearly judging *i* as the smallest, as well as judging *a* or *o* as the biggest. On the other hand, one thing that stands out as incompatible between LLM and human judgments are the ratings of vowel *e*. While humans tend to classify words containing *e* as smaller, often closer to *i*, LLMs tend to place it in a higher range.

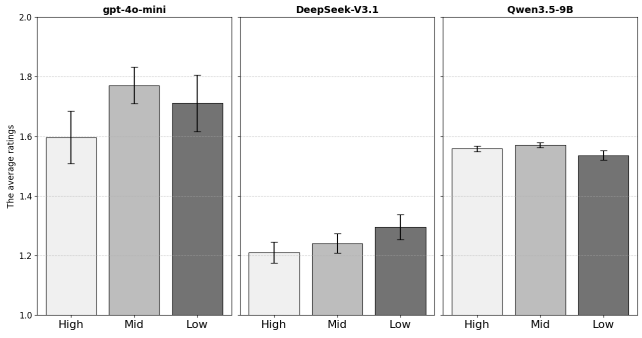
4.2 Vowel height

In the human study, the general trend revealed that the lower the vowel the higher size ratings it obtained. The significant effects compliant with this trend were found between mid and high vowels for Chinese and Japanese, as well as between low and mid vowels in Chinese and English. The only significant reversal was found in Japanese speakers that judged low vowels as smaller than mid vowels.

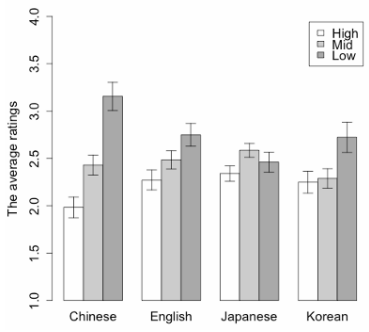
For the scale of 1 to 4, post hoc analyses of height on linear mixed effects models run for each input format type revealed no significant contrasts. However, when looking at each LLM separately, mid vowels evoked significantly larger judgments than high vowels in gpt-4o-mini ($t(113) = -3.23$, adjusted $p < 0.01$). DeepSeek-V3.1 showed no significant differences between any two levels of height. Contrastingly, Qwen3.5-9B showed a significant reversal compared to human judgments. It significantly judged high vowels as larger than low vowels ($t(113) = 4.58$, adjusted $p < 0.0001$) as well as mid vowels as larger than low vowels ($t(113) = 5.01$, adjusted $p < 0.0001$).



(a) Model size judgments on a scale of 1 to 4.



(b) Model size judgments on a scale of 1 to 2.



(c) Human size judgments on a scale of 1 to 4. Figure reproduced from Shinohara and Kawahara (2010).

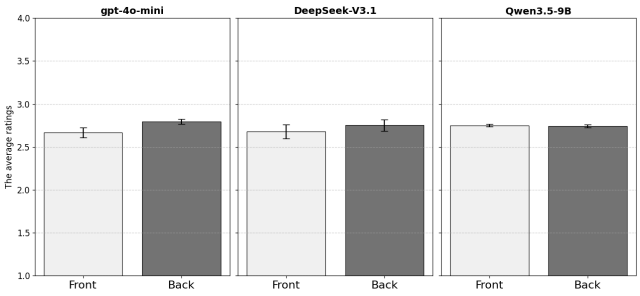
Figure 5: Average ratings across words in all input format types containing high, mid, or low vowels.

In contrast, more significant effects were found when LLMs were asked to judge words on a scale of 1 to 2. Analyses considering all three LLMs across different input format types found that mid vowels were judged as larger than high vowels both for *word* ($t(113) = -3.66$, adjusted $p < 0.01$) and *IPA* ($t(113) = -2.80$, adjusted $p < 0.05$) formats. Meanwhile, no significant effects for height judgments were found for *spaced IPA* format. Model-specific analyses revealed that gpt-4o-mini found mid vowels significantly larger than both high vowels ($t(113) = -3.88$, adjusted $p < 0.001$) as well as low vowels ($t(113) = -3.18$, adjusted $p < 0.01$). DeepSeek-V3.1 judged low vowels as significantly larger than high vowels ($t(113) = -2.38$, adjusted $p < 0.05$), while Qwen3.5-9B reversed the trend and judged low vowels as smaller than both high vowels ($t(113) = 2.61$, adjusted $p < 0.05$) and mid vowels ($t(113) = 3.92$, adjusted $p < 0.001$).

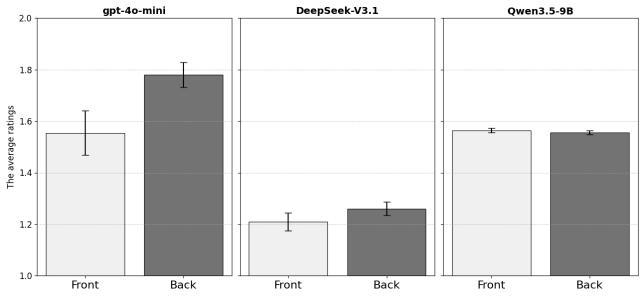
4.3 Vowel backness

Humans across all four languages judged words containing front vowels as significantly smaller than words containing back vowels. The same significant effect was found in models of all three input format types on a scale of 1 to 2 (*word*, $t(113) = -3.06, p < 0.01$; *IPA*, $t(113) = -4.04, p < 0.001$; *spaced IPA*, $t(113) = -1.98, p < 0.05$), and, surprisingly, only in the model analyzing *spaced IPA* words across LLMs on a scale of 1 to 4 ($t(113) = -2.90, p < 0.01$).

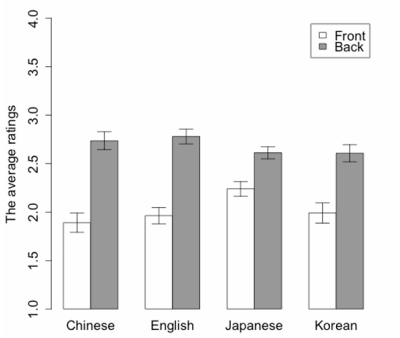
Furthermore, LLM-specific statistical analyses showed that only for gpt-4o-mini front vowels evoked significantly smaller size judgments than back vowels, both for 1 to 4 scale ($t(113) =$



(a) Model size judgments on a scale of 1 to 4.



(b) Model size judgments on a scale of 1 to 2.



(c) Human size judgments on a scale of 1 to 4. Figure reproduced from Shinohara and Kawahara (2010).

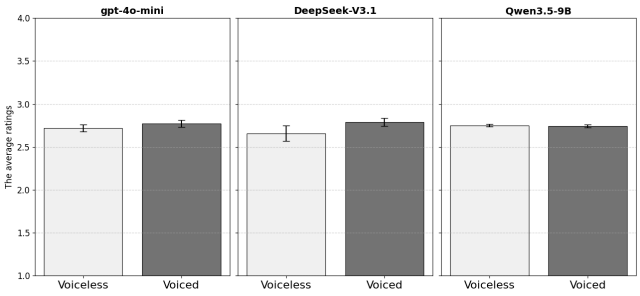
Figure 6: Average ratings across words in all input format types containing either front or back vowels.

$-4.76, p < 0.001$) and 1 to 2 scale ($t(113) = -5.78, p < 0.001$). DeepSeek-V3.1 and Qwen3.5-9B showed no significant differences between front and back vowel judgment on either size scales.

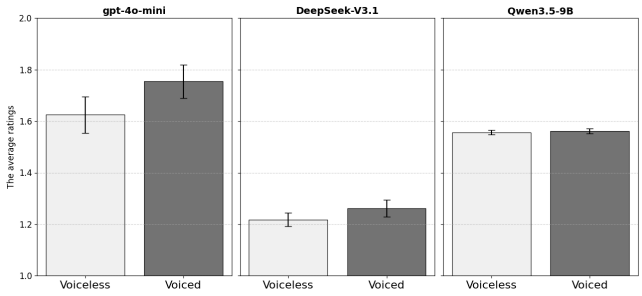
4.4 Obstruent voicing

In Chinese, English and Japanese, words containing voiceless obstruents evoked significantly smaller size judgments than words containing voiced obstruents. The same trend was found in LLM size judgments. On a scale of 1 to 4, models based on 1 judged voiceless obstruents as smaller than voiced obstruents for *word* ($t(113) = -2.05, p < 0.05$) and *IPA* ($t(113) = -2.12, p < 0.05$) format types. For *spaced IPA* no significant effect was found. Similarly, on a scale of 1 to 2 the same significant effect was found for *IPA* ($t(113) = -1.98, p < 0.05$) and *spaced IPA* ($t(113) = -3.65, p < 0.001$) format types. This time, no significant differences were observed for *word* input type.

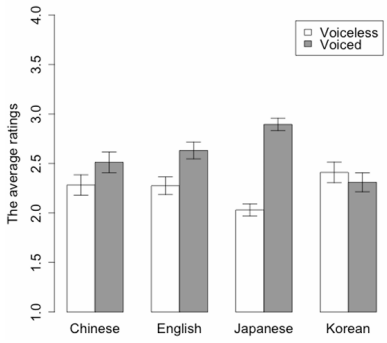
In LLM-specific analyses the trend was sustained and found significant for gpt-4o-mini and DeepSeek-V3.1 both for the scale of 1 to 4 (gpt-4o-mini, $t(113) = -2.18, p < 0.05$; DeepSeek-V3.1, $t(113) = -2.68, p < 0.01$) and the scale of 1 to 2 (gpt-4o-mini, $t(113) = -3.22, p < 0.01$; DeepSeek-V3.1, $t(113) = -2.22, p < 0.05$). On the other hand, no significant differences were found for Qwen3.5-9B.



(a) Model size judgments on a scale of 1 to 4.



(b) Model size judgments on a scale of 1 to 2.



(c) Human size judgments on a scale of 1 to 4. Figure reproduced from Shinohara and Kawahara (2010).

Figure 7: Average ratings across words in all input format types containing either voiceless or voiced obstruents.

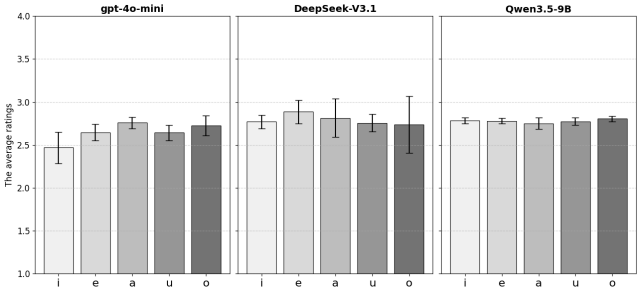
4.5 Input format type analysis

Further analysis into decisions of each LLM presented with various input format types revealed very nuanced yet important details. Starting with judgments on a scale of 1 to 4 (**Figure 8**), 95% confidence intervals of size ratings tended to be bigger in the *word* and *IPA* formats than in the *spaced IPA* format. This heavily influenced where the most significant effects were found.

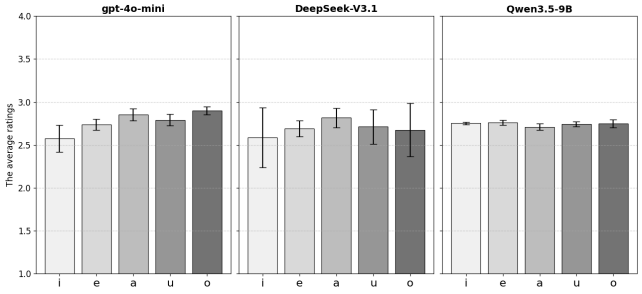
For DeepSeek-V3.1, large error bars in 8a and 8b caused no significant effects to be found in judgments of the words in these two format types. However, prompting DeepSeek-V3.1 with words in *spaced IPA* format stabilized its judgments, narrowed the error bars, and started shifting ratings towards statistical significance(front vowels smaller than back vowels ($t(35) = -1.82, p < 0.1$); high vowels smaller than both low vowels ($t(35) = -2.9, p < 0.1$) and mid vowels ($t(35) = -2.35, p < 0.1$)).

While gpt-4o-mini already showed some significant results in *word* format (front vowels smaller than back vowels ($t(35) = -2.11, p < 0.05$)), the statistical significance increased quite substantially for *IPA* format (front vowels smaller than back vowels ($t(35) = -4.21, p < 0.001$); high vowels smaller than mid vowels ($t(35) = -3.05, p < 0.05$)) as well as *spaced IPA* format (front vowels smaller than back vowels ($t(35) = -3.18, p < 0.01$); mid vowels smaller than both low vowels ($t(35) = -2.64, p < 0.05$)). Interestingly, in *spaced IPA* format words with high vowels were judged by gpt-4o-mini as larger than those containing low vowels ($t(35) = 2.78, p < 0.05$), which is a reversed trend compared to most human judgments. This shows that despite the judgments appearing visually more distinct in the *word* format, more statistically significant results for

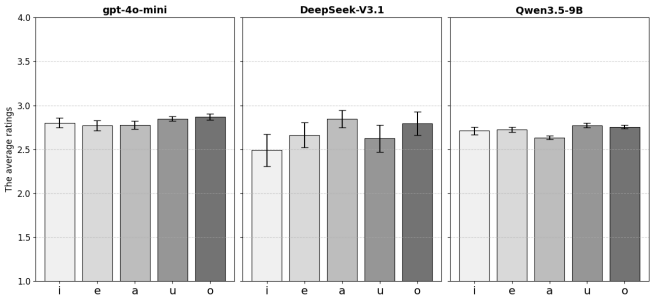
gpt-4o-mini were found in *spaced IPA* format that had considerably smaller error bars.



(a) Model size judgments in *word* input format.



(b) Model size judgments in *IPA* input format.



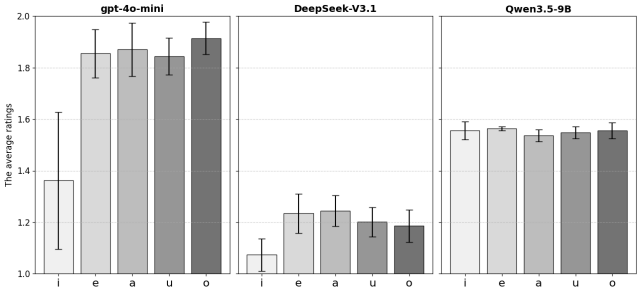
(c) Model size judgments in *spaced IPA* input format.

Figure 8: Average ratings on a scale of 1 to 4 across words in different input format types containing one of the five vowels.

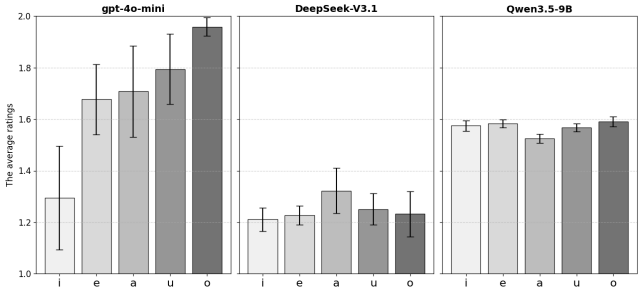
On the contrary, Qwen3.5-9B judgments appeared more visually distinct in *spaced IPA* format compared to *word* or *IPA* formats. Since the error bars remained relatively small and similar across different input format types, the only statistically significant results were found for the *spaced IPA* format (front vowels smaller than back vowels ($t(35) = -3.21, p < 0.01$); high vowels larger than low vowels ($t(35) = 6.80, p < 0.0001$) and low vowels smaller than mid vowels ($t(35) = -6.69, p < 0.0001$)). These results show that the strong reversal in height judgments for Qwen3.5-9B compared to human judgments was primarily driven by words in *spaced IPA* format.

Surprisingly, input format types had a completely different effect on LLM size judgments on a scale of 1 to 2 (**Figure 9**) compared to the scale of 1 to 4. Here, *spaced IPA* format did not cause such a noticeable decrease of error bars as it did on a scale of 1 to 4. Instead, in some instances, it made the judgments even more uncertain. For example, *spaced IPA* format made Qwen3.5-9B judgments more uniform and caused an increase in error bars, leading to no significant effects being found when judging words in this format. While the *word* format also did not evoke any statistically significant results for Qwen3.5-9B, in the *IPA* format the reversed trend of height of the vowels affecting size showed up as highly significant (high vowels larger than low vowels ($t(35) = 3.92, p < 0.01$) and low vowels smaller than mid vowels ($t(35) = -5.25, p < 0.0001$)).

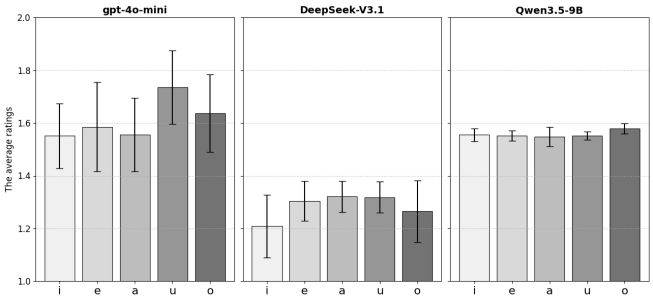
Spaced IPA format made gpt-4o-mini judgments, yet again, more uniform while failing to reduce the error bars (9c), resulting in only one significant effect found (voiceless smaller than voiced ($t(35) = -3.13, p < 0.01$)). Despite large error bars being found in other formats as well, but because of massive differences between vowel size judgments, more statistically significant results were found for the *word* input format type (9a) (front vowels smaller than back vowels



(a) Model size judgments in *word* input format.



(b) Model size judgments in *IPA* input format.



(c) Model size judgments in *spaced IPA* input format.

Figure 9: Average ratings on a scale of 1 to 2 across words in different input format types containing one of the five vowels.

($t(35) = -3.46, p < 0.01$); high vowels smaller than mid vowels ($t(35) = -3.60, p < 0.01$)) as well as the *IPA* format (9b) (front vowels smaller than back vowels ($t(35) = -5.21, p < 0.001$); high vowels smaller than mid vowels ($t(35) = -3.64, p < 0.01$) and mid vowels smaller than low vowels ($t(35) = -3.07, p < 0.05$)).

Lastly, DeepSeek-V3.1 showed no significant effects for the *word* input type, while both for *IPA* and *spaced IPA* formats words containing voiceless obstruents were judged as significantly smaller than those containing voiced obstruents (*IPA*, $t(35) = -2.18, p < 0.05$; *spaced IPA*, $t(35) = -2.08, p < 0.05$).

Altogether, prompting models with words in different input formats caused various reactions depending on the scale of size judgments as well as the LLM itself. On a scale of 1 to 4, *spaced IPA* format seemed to extract the differences between phonetic features much more effectively than the two other formats. Despite visually judging various vowels less distinctly for some models (gpt-4o-mini) and more distinctly for other (Qwen3.5-9B), the format managed to make LLM judgments more precise, decreasing the 95% confidence interval and finding more statistically significant effects. On the other hand, for size judgments on a scale of 1 to 2, *IPA* format managed to keep model vowel size judgments distinct so despite relatively large error bars, the significant effects could be best observed in this format.

5 Discussion

The presented analysis shows that all three LLMs exhibit some degree of sound-symbolic associations. Gpt-4o-mini consistently judged words containing mid vowels to be the largest compared to other vowel heights, words containing front vowels to be smaller than those containing back vowels, and words containing voiceless obstruents to be smaller than those containing voiced obstruents. DeepSeek-V3.1 found words containing low vowels larger than those with high vowels and words containing voiceless obstruents smaller than those containing voiced obstruents. Lastly, Qwen3.5-9B, in general, across the three input format types judged the words significantly different only based on vowel height, judging words containing low vowels as smaller than those containing mid or hight vowels.

Overall, humans from the study by Shinohara and Kawahara (2010) judged words such that they rated front vowels as smaller than back vowels, voiceless obstruents as smaller than voiced obstruents and the low vowels larger than mid vowels, which were in turn larger than high vowels. That being said, gpt-4o-mini turned out to reflect human size sound-symbolic judgments most accurately, with the only inconsistency in mid vowels being judged as larger than low vowels. DeepSeek-V3.1 also partly demonstrated the same size sound-symbolic associations as human participants. In contrast, Qwen3.5-9B showed a highly significant reversal of the vowel height judgment trend compared to humans.

Interestingly, specific LLMs also tend to follow trends of certain language speakers more prominently than others. For example, gpt-4o-mini can be seen consistently judging *i* as the smallest while *o* as the largest (4b), similarly to judgments of English and Japanese participants (4d). Meanwhile, DeepSeek-V3.1 has a tendency of judging *a* as considerably larger than any other vowel, similarly to Chinese participants. This is slightly surprising as both the models are known to be trained on balanced data in a variety of different languages. Nonetheless, this observation, although small, could suggest that even multilingual models may be susceptible to very nuanced biases hidden within their dominant languages.

5.1 Effects of models on size sound-symbolic associations

As described above, gpt-4o-mini reproduces human size sound-symbolic associations most precisely out of all the models, despite its scaled-down architecture. There could be a number of reasons associated with such a precise performance. One of them could be the fact that gpt-4o-mini is natively multimodal, so while it can function as a pure LLM, it has been trained on data in multiple modalities, also integrated together. Since sound symbolism is a cross-modal phenomenon, these biases could have been learned and emphasized in training, transferring the associations to textual tasks.

However, Qwen3.5-9B which is also natively a multimodal model, does not completely reflect human size sound-symbolic associations. A part of the explanation as to why that happens could be in the size of the model as Qwen3.5-9B has only 9B parameters. However, more reasons could play a role in its failure to follow human sound-symbolic biases.

Nevertheless, multimodal training may not be necessary to obtain human-like size sound symbolism in LLMs, which is proven by DeepSeek-V3.1. Its judgments are precise, and while not all phonetic features evoke significant differences in size judgments, DeepSeek-V3.1 demonstrates that human size sound-symbolic biases can be extracted purely from text. In this case the size of

the model seems to matter, DeepSeek-V3.1 has been trained on massive amounts of textual data which likely allows it to extract the associations without ever having access to other modalities.

5.2 Effects of size judgment scales on size sound-symbolic associations

Analysis into using different scales when prompting LLMs for size judgments showed the importance of taking into account model's biases and testing model behavior to ensure compliance with the task. While humans tend to display a certain level of survey response bias when asked to make scaled judgments, LLMs also display such biases, often avoiding extreme ends of the scale (Rupprecht et al., 2025). Taking that into account, when trying to replicate human-like magnitude sound-symbolic associations it is crucial to ask for size judgments on a carefully chosen scale to capture the most of these mappings and minimize the impact of central tendency of LLMs on the results. Findings in this paper have shown that using the original scale of 1 to 4 makes LLM vowel size judgments relatively homogeneous. In spite of the low visual distinction between the vowels, judgments on this scale are extremely precise, which translates to highly significant results.

On the other hand, changing the scale into forced choice (1 or 2) made the models' vowel size judgments more divergent. However, at the same time this introduced a much higher uncertainty, which made the differences between magnitude judgments of various phonetic features much less significant. It seems that forcing models into making a strict choice between two polarities, small and large, made their decisions much more unstable. This could insinuate that size sound symbolism is a very nuanced phenomenon that becomes extremely susceptible to high variance when being judged on a small scale. Perhaps the subtle associations of phonetic features and size can only be made confidently when presented on a complex-enough scale.

Although this could lead to thinking that the wider scale would provide a better insight into the nuanced sound-symbolic associations, there seem to be limitations on that end as well. For example, the introduction of a larger scale(1 to 8), which was supposed to achieve clearer sound-symbolic trends within the middle values of the scale, did not yield attractive results. The wider scale seemed to confuse some of the models, specifically DeepSeek-V3.1 which started to ignore prompt commands and return unwanted tokens. While other models may have successfully mapped size sound-symbolic associations on that scale, and it was just one model that started misbehaving, the results were clearly compromised and unsuitable for further analysis. This again points to the importance of choosing a correct scale when investigating magnitude sound symbolism in LLMs.

5.3 Effects of stimulus format types on size sound-symbolic associations

A closer look at the effects of input format types on size sound-symbolic judgments in LLMs revealed an interesting trend that could suggest where the origin of this phenomenon in these models. On both scales, 1-2 and 1-4, the *word* format was not the one that yielded the most statistically significant results. Instead, on a scale of 1 to 4, the *spaced IPA* format made the models' judgments much more precise, and despite also making them more homogeneous, the reduced error bars allowed the effects to be much more statistically significant. On the other hand, on a scale of 1 to 2, the *IPA* format maintained the diverse vowel size judgments made by models while also relatively limiting uncertainty, which again caused more statistically significant results to be visible. This brings an important point to the conversation about the emergence of magnitude sound symbolism phenomenon within LLMs. The fact that *word* does not yield the highest statistically

significant results could suggest that the phonetic associations are not directly 'remembered' from the training data. Instead, making models pay special attention to phonetic representations of words through the use of both IPA formats could be causing the sound symbolism phenomenon to emerge more. This is an important discovery showing that LLMs may be more sensitive to the actual phonetic mappings than initially thought, instead of heavily relying on pre-learned data.

6 Limitations and future directions

While the results of this study provide valuable insight into the nature of magnitude sound symbolism phenomenon and its emergence in LLMs, there are some limitations that need to be considered. First, it is impossible to know to what extent the models may have memorized and learned human-like magnitude sound-symbolic associations directly from training data. The topic of sound symbolism is widely described and tested in both the academic and the general literature on which the models may have been trained. While introducing IPA format types may help in decreasing models' reliance on pre-learned words, it does not guarantee complete zero-shot generalization. However, the highly mixed results described in this paper suggest that the models do not simply memorize and retrieve the information they have seen before. Quite the opposite seems to be true: the *word* input format type does not yield the most significant results in most cases. Rather, the *IPA* and *spaced IPA* formats, which force the models to pay attention to individual phonetic structures, show stronger statistical significance and narrower confidence intervals. Moreover, in some instances models show magnitude sound-symbolic associations but with the reversed trend compared to humans, for example, vowel height judgments of Qwen3.5-9B. This suggests that the associations are not blindly learned and repeated, but rather processed in real time based on small phonetic nuances.

Moreover, the experiment was conducted on only three LLMs. This was due to limited model availability and resources. Despite the model selection being conducted carefully to ensure a wide and varied representation of architectures, evaluating only three models poses restrictions to the true generalizability of the findings to all currently available LLMs. Moreover, details about the architecture and training of gpt-4o-mini model used in this study are not available to the public. This makes it more difficult to draw conclusions about the possible effects that model architectures and training may have on the studied phenomenon.

Furthermore, as revealed in the analysis, prompting conditions seem to significantly impact the way magnitude sound-symbolic associations show up in LLMs. This again raises concerns about the generalizability as well as reliability of the results. Thus, it is important to note that while human-like size sound-symbolic biases can emerge in LLMs, their presence and strength is highly sensitive to the conditions of the testing environment. These limitations emphasize the need for further research into magnitude sound symbolism within a more extensive selection of LLMs and across various testing conditions to strengthen the robustness of the findings.

7 Conclusion

This paper explored how chosen Large Language Models respond to a task that reveals biases of human embedded cognition. It investigated magnitude sound-symbolic associations, which are mappings between phonetic features, such as vowel height, and imagined sizes. While this

phenomenon is cross-modal in nature and in humans it emerges from auditory grounded experience, this study explored whether size sound-symbolic mappings remain strong enough in text alone to be displayed by LLMs.

The experiment was conducted on three LLMs (gpt-4o-mini, DeepSeek-V3.1, and Qwen3.5-9B) and evaluated against a cross-linguistic human study. The findings revealed clear answers to **RQ1** and **RQ2**. Magnitude sound-symbolic associations can indeed show up in computational models, although their agreement with human judgments varies depending on the model. Gpt-4o-mini showed the highest similarities with human trends. It accurately mirrored the significant effects of vowel backness and obstruent voicing on size judgments, and displayed striking preference of judging the vowel *i* as the smallest. DeepSeek-V3.1 also exhibited significant size sound-symbolic effects, successfully reflecting part of human biases, specifically those of obstruent voicing and low versus high vowel trends. In contrast, the smallest of the models, Qwen3.5-9B, also displayed some highly significant sound-symbolic associations; however, they were a complete reversal of human vowel-height trends. This showed that while all models seem to exhibit magnitude sound-symbolic biases, not all of them are aligned to the human judgments.

Importantly, not only did the model architecture have an impact on size judgment trends, but also other factors played an important role in shaping LLMs' responses. While the forced binary choice (1-2 scale) visually polarized the differences between vowels, it also introduced high uncertainty in the decisions. Meanwhile, a wider scale of 1 to 4 managed to capture significantly different, very precise, yet low in variance vowel size judgments. This revealed that a more complex scale was better at precisely extracting the nuanced nature of magnitude sound symbolism.

Finally, addressing **RQ3**, the effects of input formats used when prompting the models provide some valuable insights into the emergence of size sound symbolism. The experiment revealed that the *word* format did not yield the most and highest statistically significant results. Instead, it was the *IPA* and *spaced IPA* formats that drove most of the effects. This suggests that memorized tokens from the training data are not the primary driving forces that make the LLMs form these associations. Rather, models are more sensitive to phonetic features emphasized in IPA formats. Moreover, the findings suggest that no access to auditory information is necessary for the magnitude sound-symbolic biases to emerge. As presented, even the purely text-only model DeepSeek-V3.1 was able to extract the phonetic mappings accurately. This implies that size sound-symbolic associations can be deeply rooted into statistical distributional patterns of human language, from which LLMs extract information. While it is certain that sound plays an important role in shaping magnitude sound symbolism in humans, this research showed that it is possible to reconstruct these human biases through the textual data alone.

References

Qwen/Qwen3.5-9B · Hugging Face — huggingface.co. https://huggingface.co/{Q}wen/{Q}wen3.5-9{B}. [Accessed 21-05-2026].

Morris Alper and Hadar Averbuch-Elor. Kiki or bouba? sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, 36:78347–78359, 2023.

Brent Berlin. The first congress of ethnozoological nomenclature. *Journal of the Royal Anthropological Institute*, 12:S23–S44, 2006.

Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823, 2016.

Christine Cuskley, Mark Dingemanse, Simon Kirby, and Tessa M Van Leeuwen. Cross-modal associations and synesthesia: Categorical perception and structure in vowel–color mappings in a large online sample. *Behavior research methods*, 51(4):1651–1675, 2019.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841), 2022.

Ferdinand De Saussure et al. Nature of the linguistic sign. *Course in general linguistics*, 1:65–70, 1916.

Mark Dingemanse, Damián E Blasi, Gary Lupyan, Morten H Christiansen, and Padraic Monaghan. Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, 19(10): 603–615, 2015.

Olga Fischer, Kimi Akita, and Pamela Perniss. *The Oxford Handbook of Iconicity in Language*. Oxford University Press, 2026.

Patrice L French. Toward an explanation of phonetic symbolism. *Word*, 28(3):305–322, 1977.

Alberto Gallace and Charles Spence. Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & psychophysics*, 68(7):1191–1203, 2006.

Charles F Hockett and Charles D Hockett. The origin of speech. *Scientific American*, 203(3):88–97, 1960.

Mutsumi Imai, Sotaro Kita, Miho Nagumo, and Hiroyuki Okada. Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65, 2008.

Jinhong Jeong, Sunghyun Lee, Jaeyoung Lee, Seonah Han, and Youngjae Yu. Do language models associate sound with meaning? a multimodal study of sound symbolism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31247–31255, 2026.

Wolfgang Köhler. *Gestalt Psychology*. Liveright, New York, NY, 1929.

Tom Kouwenhoven, Kiana Shahrasbi, and Tessa Verhoef. Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 76452–76479. Curran Associates, Inc., 2025.

Wolfgang Köhler. *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. Horace Liveright, New York, 2nd edition, 1947.

Mallorie Leinenger. Phonological coding during reading. *Psychological bulletin*, 140(6):1534, 2014.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Tyler Loakman, Yucheng Li, and Chenghua Lin. With ears to see and eyes to hear: Sound symbolism experiments with multimodal large language models. *arXiv preprint arXiv:2409.14917*, 2024.

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

Lawrence E Marks. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of experimental psychology: Human perception and performance*, 13(3):384, 1987.

Daphne Maurer, Thanujeni Pathman, and Catherine J Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental science*, 9(3):316–322, 2006.

Stanley S Newman. Further experiments in phonetic symbolism. *The American Journal of Psychology*, 45(1):53–75, 1933.

Alan KS Nielsen and Mark Dingemanse. Iconicity in word learning and beyond: A critical review. *Language and speech*, 64(1):52–72, 2021.

John J Ohala. The frequency code underlies the sound-symbolic use of voice pitch. *Sound symbolism*, pages 325–347, 1995.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence — openai, Jul 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Cesare V Parise and Charles Spence. Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, 220(3):319–333, 2012.

Melissa A Preziosi and Jennifer H Coane. Remembering that big things sound big: Sound symbolism and associative memory. *Cognitive Research: Principles and Implications*, 2(1):10, 2017.

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

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. Prompt perturbations reveal human-like biases in llm survey responses. *arXiv preprint arXiv:2507.07188*, 2025.

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

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

David M Sidhu and Penny M Pexman. Five mechanisms of sound symbolic association. *Psychonomic bulletin & review*, 25(5):1619–1643, 2018.

Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.

Patrick D Thompson and Zachary Estes. Sound symbolic naming of novel objects is a graded function. *Quarterly journal of experimental psychology*, 64(12):2392–2404, 2011.

Tessa Verhoef, Kiana Shahrasbi, and Tom Kouwenhoven. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, 2024.

Vincent Walsh. A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in cognitive sciences*, 7(11):483–488, 2003.