With Ears to See and Eyes to Hear: Sound Symbolism Experiments with Multimodal Large Language Models

Tyler Loakman[1], Yucheng Li[2] and Chenghua Lin[1,3]*
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, University of Surrey, UK
[3]Department of Computer Science, The University of Manchester, UK
tcloakman1@sheffield.ac.uk
yucheng.li@surrey.ac.uk chenghua.lin@manchester.ac.uk

Abstract

Recently, Large Language Models (LLMs) and Vision Language Models (VLMs) have demonstrated aptitude as potential substitutes for human participants in experiments testing psycholinguistic phenomena. However, an understudied question is to what extent models that only have access to vision and text modalities are able to implicitly understand sound-based phenomena via abstract reasoning from orthography and imagery alone. To investigate this, we analyse the ability of VLMs and LLMs to demonstrate sound symbolism (i.e., to recognise a non-arbitrary link between sounds and concepts) as well as their ability to "hear" via the interplay of the language and vision modules of open and closed-source multimodal models. We perform multiple experiments, including replicating the classic Kiki-Bouba and Mil-Mal shape and magnitude symbolism tasks and comparing human judgements of linguistic iconicity with that of LLMs. Our results show that VLMs demonstrate varying levels of agreement with human labels, and more task information may be required for VLMs versus their human counterparts for *in silico* experimentation. We additionally see through higher maximum agreement levels that Magnitude Symbolism is an easier pattern for VLMs to identify than Shape Symbolism, and that an understanding of linguistic iconicity is highly dependent on model size.

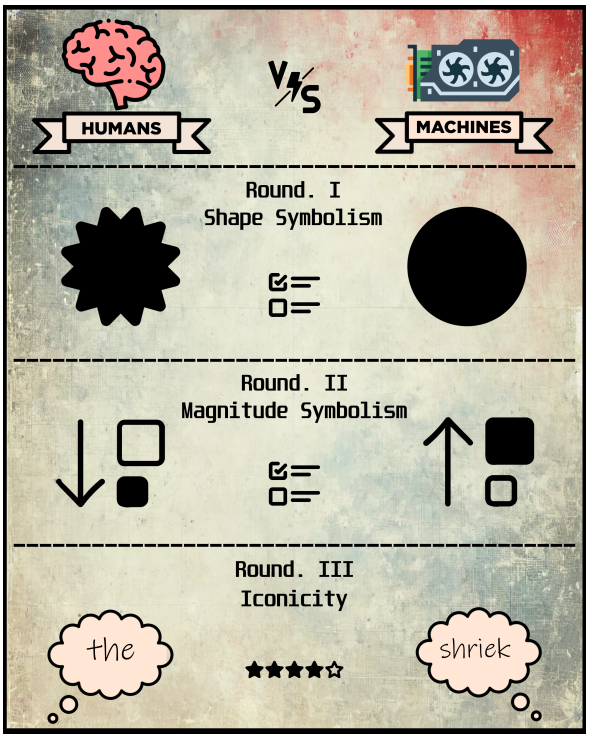

Figure 1: Illustration of the 3 main experiments we perform. Firstly, Shape Symbolism is a binary choice between two pseudowords to best describe an object that is spiky or rounded. Magnitude Symbolism involves a binary choice between two pseudowords to best describe an object that is small or large. Finally, Iconicity involves rating the perceived iconicity of words, or how much their written/phonetic form is representative of what they describe.

1 Introduction

Sound symbolism refers to a perceived similarity between speech sounds and the conceptual meanings of the words they comprise. Evidence of this can be found in linguistic devices such as onomatopoeia (e.g., "bang", "shriek", and "bellow"), where a word imitates the concept it describes via its phonetic form. The ability of Large Language Models (LLMs) and Vision Language Models (VLMs) to reflect a sense of sound symbolism would therefore suggest that these models are capable of acquiring "phonetic" knowledge indirectly through only the written orthographic form of a language via patterns of grapheme combinations (Loakman et al., 2024, 2023) and meta-level textual discussion of sound in training data, which has implications for the potential future use of LLM/VLMs in perceptual studies usually reserved for humans (Jain et al., 2023; Dillion et al., 2023; Aher et al., 2023). In this work, we further explore the capability of LLMs/VLMs to demonstrate human-like characteristics in a range of psycholinguistic perceptual

*Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2849–2867
November 12-16, 2024

tests investigating 3 main areas of sound symbolism: (1) *Shape Symbolism* (i.e., the Kiki-Bouba effect, Ramachandran and Hubbard, 2001), where a forced choice must be made between two pseudowords as to which is the most appropriate to describe shapes and entities that are spiky or rounded; (2) *Magnitude Symbolism* (i.e., the Mil-Mal Effect, Sapir, 1929), a similar test to (1), but where the entities are small or large (rather than spiky or rounded); and (3) *Iconicity Rating* (Winter et al., 2023), where LLMs are asked to rate a series of English words on their perceived "iconicity" (i.e., to what extent a word's form is perceived to be analogous to the concept or entity it describes). We illustrate these experiments in Figure 1.

By extending these experiments to LLMs/VLMs, our study aims to shed light on the processes underlying multimodal perception in language models.[1] Moreover, the presence of sound symbolism in such models could inform the development of more effective natural language processing algorithms, aiding tasks such as sentiment analysis, emotion recognition, and content generation that take into account more abstract layers of human reasoning and perception such as abstract connotations between words rather than semantics alone (Manzoor et al., 2023). An understanding of sound symbolism also has the potential to have a profound effect on creative generation, including language forms such as poetry and narratives (and their accompanying illustrations). Additionally, sound symbolism is a prevalent strategy used in marketing products to create desirable associations in potential customers, and LLMs capable of understanding this phenomenon could be used as pilot testing before using focus groups to reduce time and monetary costs (Ketron and Spears, 2021; Motoki et al., 2020; Spence, 2012).

We summarise our main contributions as follows:[2]

- We perform replications of the classic psycholinguistic Kiki-Bouba Shape Symbolism and Mil-Mal Magnitude Symbolism studies with a range of open and closed-source VLMs to investigate if they understand the association between speech sounds/orthographic forms and the characteristics of entities.
- We perform an in-depth analysis of the ability of a range of closed and open-source LLMs to demonstrate an understanding of linguistic iconicity by comparing judgements to an existing large-scale dataset of human ratings.
- We provide a discussion of the potential sources of sound symbolism abilities in LLM/VLMs and potential future approaches to bolstering these abilities, in addition to the implications of doing so in §6.

2 Related Works

The work of early linguists, such as Saussure, touched upon the topic of whether or not the link between the "sign" (i.e., a word) and the "signified" (i.e., the entity/concept to which the sign relates) is arbitrary, with there being nothing more "boat"-esque about the word "boat" than any other combination of phonotactically legal sounds (de Saussure and Baskin, 2011). However, there are many types of language where this association is seen to be *non*-arbitrary such as in the onomatopoeia commonly used in literary works (e.g., "bang" for a loud noise, or "shriek" for a high-pitched wail), where the phonetic realisation mirrors the concept it describes. These phenomena as a whole are known as sound symbolism, where there is thought to be a non-arbitrary link between the sign and the signified, in contrast with the popular stance of early linguistics.

Outside of onomatopoeia, sound symbolism is believed to have a range of effects on human perception, including applying to nonsense pseudowords. For example, even if a word was not created to explicitly denote a known concept or entity (and therefore has no true *denotative* meaning), it is nevertheless able to manifest a *connotative* meaning in the mind of the reader based on its phonological representation and/or phonetic realisation. The first identification of these patterns is frequently attributed to Usnadze (1924), who gave 10 participants a series of pseudowords alongside drawings and found a higher-than-chance level of agreement between evaluators for which nonsense word best described which drawing. Perhaps the most famous example of this is in the Kiki-Bouba effect (Sidhu et al., 2021; Ramachandran and Hubbard, 2001; Köhler, 1929) which concerns the allocation of the name "Kiki" to sharp, hard-edged entities, and "Bouba" to more soft and round-edged entities (which in the original works consisted exclusively of 2D shapes). A similar relationship has also been noted between the words "Mil" and "Mal", where the changing vowel in the phonological min-

[1] Our paper title is inspired by the Sleeping With Sirens album of the same name: https://en.wikipedia.org/wiki/With_Ears_to_See_and_Eyes_to_Hear.

[2] We release our code and resources on GitHub: https://github.com/tylerL404/WETSAETH/.

imal pairs[3] has a relationship to perceived size, in a phenomenon known as magnitude symbolism (where vowels with higher frequency content are associated with smaller entities due to the relationship between vocal tract length and vocal productions) (Sapir, 1929). Extensive research has been performed in the area of sound symbolism, demonstrating interesting findings such as these patterns being largely language agnostic (Ćwiek et al., 2022) as well as being weaker in neurodivergent individuals (Occelli et al., 2013) and not being yet developed in very early childhood (Sidhu et al., 2023). Other research has also investigated the exact requirements and limits of the effects (Sidhu and Vigliocco, 2023; Passi and Arun, 2022; Styles and Gawne, 2017; Nielsen and Rendall, 2013).

Recently, Alper and Averbuch-Elor (2023) have investigated the ability of language models to exhibit the Kiki/Bouba effect using CLIP and Stable Diffusion by generating images from sound-symbolic prompts and provide positive evidence for this association to be present. We build upon this work in §3 by introducing a wider range of VLMs in a forced naming task for "real" entities as opposed to abstract shapes, and additionally extend this to magnitude symbolism via the Mil-Mal test in §4. We further differentiate our work by focussing on the task of assigning pseudowords to provided visual stimuli, in contrast to Alper and Averbuch-Elor (2023) who investigate the effects of different pseudowords on the outputs of image generation models.

Additionally, in recent times, large-scale efforts have been made to collect ratings of linguistic iconicity (i.e., the level of symbolism a particular word has), with Winter et al. (2023) collecting ratings of over 14k English words. Some effort has been made to analyse whether or not similar ratings would be assigned by an LLM, where Trott (2024) used GPT-4 and reports a moderate positive correlation across ratings. We build upon this work in §5 by introducing a wider range of VLMs, including open-source alternatives. Numerous computational works in NLP have investigated other aspects of iconicity and sound symbolism, with Abramova and Fernández (2016) investigating the word embeddings of different aspects of morphology in relation to symbolism (see also Yamshchikov et al., 2019; Liu et al., 2018). Additionally, Sabbatino et al. (2022) investigated the emotional intensity of nonsense words using NLP methods to determine which phoneme combinations were most responsible.

[3]A *Minimal Pair* refers to a pair of words that differ only in one phonological segment, such as "cat" /kat/ versus "bat" /bat/.

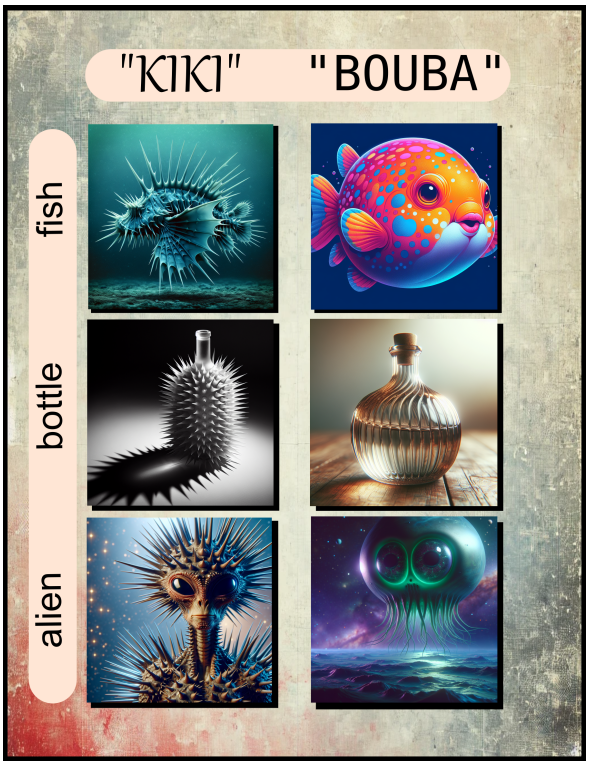

Figure 2: Examples of "Kiki"-style (spiky) and "Bouba"-style (rounded) generations with DALL-E 3. In total, 50 images were generated, with 25 per condition (the entities remaining constant). The ground truth is taken as the majority human vote.

Several works in similar areas have also exemplified the ability of LLMs to demonstrate perceptual behaviour akin to humans and the potential for these models to replace human participants in pilot studies, as well as facilitating the scaling of evaluation *in-silico* (Jain et al., 2023; Aher et al., 2023; Dillion et al., 2023; Ramezani and Xu, 2023; Coda-Forno et al., 2023).

3 Shape Symbolism

In this section, we perform a replication of the classic Kiki-Bouba Effect experiment (Ramachandran and Hubbard, 2001) using a range of multi-modal LLMs. Within the traditional set-up for the Kiki/Bouba test, human participants are presented with either a rounded soft-edged shape or a spiky sharp-edged shape and asked to assign one of two pseudowords to either. In numerous experiments (Sidhu et al., 2021; Ramachandran and Hubbard, 2001; Köhler, 1929), words such as "Bouba" and "Maluma" are preferred for the latter rounded shapes, whilst "Kiki" and "Takete" are preferred for the spiked shapes. These findings are thought to demonstrate a non-arbitrary link between particular speech sounds and the physical characteristics of the shapes to which they refer.

3.1 Methodology

Image Dataset We prompt DALL-E 3 (Betker et al., 2023) to generate a series of images pertaining to entities that are either "spiky" or "rounded". Example generations can be seen in Figure 2. We use DALL-E 3 to generate examples rather than taking existing images such as the traditional representation of Kiki-Bouba in order to reduce the effects of memorisation and exhibit finer control of the physical characteristics of the presented entities. Furthermore, this allows us to further investigate the extent to which human perception of this phenomenon extends from geometric shapes to entities in the "real" world, therefore increasing ecological validity and more closely representing how LLM/VLMs may be tasked with demonstrating sound symbolism in the real world (e.g., when naming new products for marketing or new characters in a narrative). We use the prompt "*Generate an image from the following description: [spiky/rounded + noun]*", where we generate 25 spiky examples, and 25 round examples (each noun is used once per shape condition). The full list of entities can be seen in Appendix A.1.

Pseudowords Similar to our need to generate novel imagery to better ensure the non-memorisation of our chosen VLMs, we must avoid bias from the eponymous pseudowords (i.e., "Kiki" or "Takete" for spiked concepts, or "Bouba" and "Maluma" for round concepts). As a result of this, we consult existing sound symbolism research papers and select the following pseudowords that are legal in English phonotactics and imitate the same phonetic relationship that the original terms were meant to elicit. Borrowing from Occelli et al. (2013) we take: *Kalika–Mabobe*, *Zaki-Umbu*, and *Tiki-Giba*. From Alper and Averbuch-Elor (2023) we additionally take *Kitaki-Gugagu*, *Hatiha-Bodubo* and *Penape-Gunogu*. Finally, we use the original *Kiki-Bouba* names as a point of reference for a best-case scenario where the link is explicitly learnt from mentions within the training data.

Task Setting We imitate the standard human setup for the Kiki/Bouba experiment (Ramachandran and Hubbard, 2001) and present our VLMs with the following zero-shot prompt – *"Look at the [ENTITY] in the provided image. Out of the following two options, which name would you most likely assign to the [ENTITY]: "[KIKI-WORD]" or "[BOUBA-WORD]". Respond with only your decision"*. One candidate from either name category (i.e., Kiki or Bouba) is presented as outlined previously, and [ENTITY] refers to a noun used to describe the entity we wish to be named in order to direct the LLM's attention to the correct element (i.e., the noun from the DALL-E prompt). For this experiment, we set max_tokens to 10 for the VLM responses and leave all other hyperparameters at default. We additionally provide an extended prompt we call *informed*, which prepends "*This task is related to the phenomenon of Sound Symbolism, which is a non-arbitrary relationship between the sound of a word and associations with its physical attributes*" to give the VLMs additional task knowledge. This secondary prompt scenario is used to investigate whether human-like preferences can be encouraged from the model with additional awareness of which elements of the image to focus on. We present each prompt twice, placing each pseudoword in the first or second position and then averaging the results, to mitigate positional biases in selections.

Models We use a selection of open- and closed-source VLMs, including multimodal **GPT-4** (OpenAI et al., 2023), **Gemini Pro** (Reid et al., 2024), and **LLaVA** (Liu et al., 2023). For our open-source LLaVA model, we investigate whether sound symbolism effects arise as a direct factor of model size by including the 7-, 13-, and 34-billion parameter versions. Implementation details are given in Appendix A.1.

Evaluation As our human point of comparison, we recruited 10 human evaluators with native-level English proficiency[4] via internal methods (i.e., email lists at the primary author's institution and word-of-mouth) and presented an analogous task to that which we present to the VLMs. The order of image presentation to participants is randomised to avoid order effects.

3.2 Results

Overall, in Figure 3 we see mixed results as to which model performs best, with GPT-4 showing the highest levels of agreement for the original Kiki-Bouba and the added Kalika-Mabobe, Gemini performing the best for Zaki-Umbu and Hatiha-Bodubo, and LLaVA performing the best for the remaining conditions. However, across all models, we see a general trend of low agreement with human ratings, with only a few condition/model combinations resulting in agreement above chance (50%). Regarding the introduction of the "informed" prompt (containing additional task information), we see a general increase in agreement over the Standard condition or no change in results,

[4]Evaluators were recruited in different waves following revisions. All evaluators were paid above the current UK Living Wage per hour.

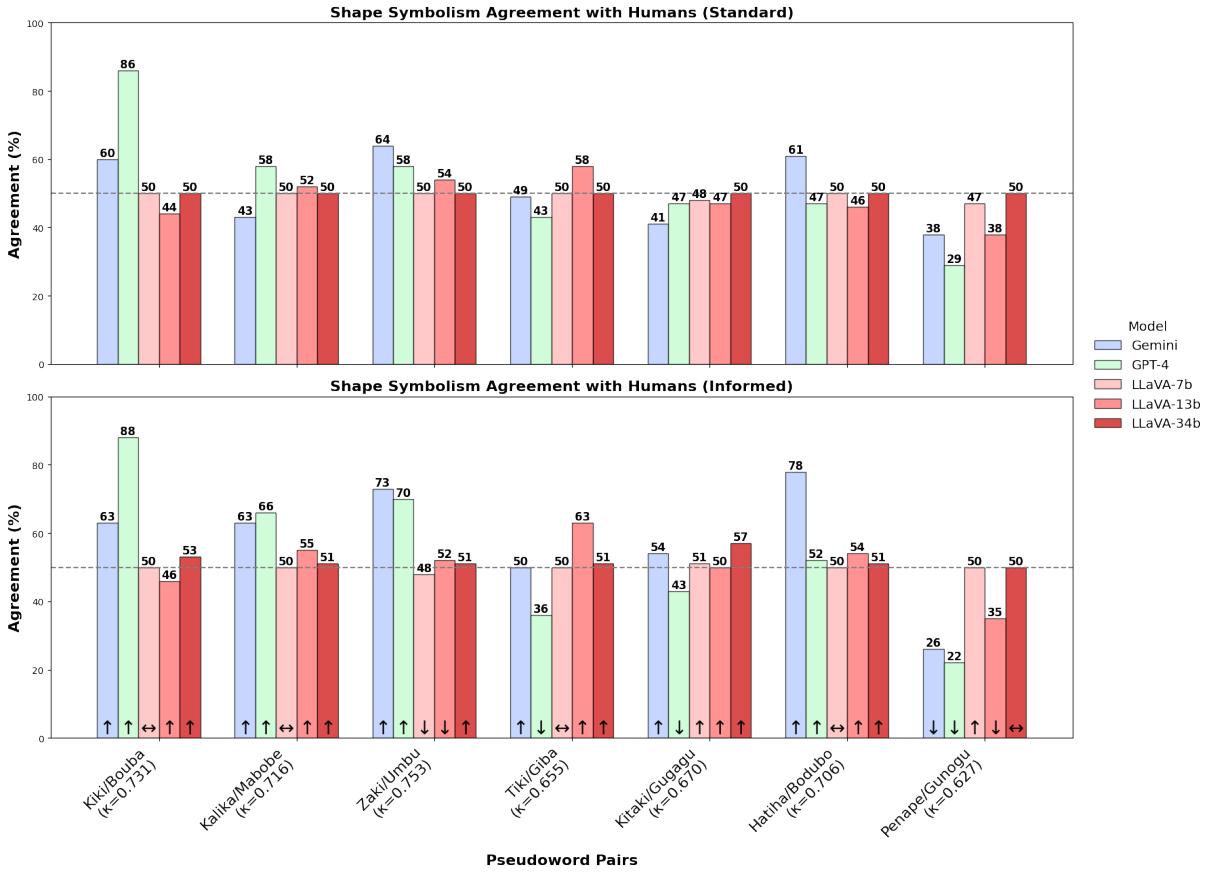

Figure 3: Results of the Shape Symbolism experiments per pseudoword pair. Fleiss' κ (Fleiss, 1971) for inter-annotator agreement between humans is presented next to each pseudoword pair. Arrows indicate the direction of agreement change from the standard prompt to the informed prompt. The dashed line represents 50%, akin to chance-level agreement. Full results in table form are presented in Table 2 within Appendix A.2. In all cases, we are comparing with the human majority vote.

indicating that once the VLMs are aware of the characteristic of interest (i.e., the shape of the entity), the VLMs are more likely to agree with human perception. However, we do see a few cases (e.g., Zaki/Umbu with LLaVA 7/13b, Kitaki/Gugagu with GPT-4, and Penape/Gugagu with Gemini, GPT-4, and LLaVA 13b) where performance decreases in the informed condition, but these are usually only minor decreases except for cases where there is already very low agreement with humans. We see varying performance across open and closed-source models. Whilst LLaVA outperforms GPT-4 and Gemini in many conditions, this is often close to chance-level agreement and may be the result of label bias, versus the closed models' systematic disagreement.[5] Finally, regarding our open-source LLaVA models at different sizes, we interestingly see that the 13B model outperforms the 7B and 34B models at times, most noticeably in Tiki-Giba. However, the largest 34b model performs best overall, in line with expectation (though by a small margin). Additionally, the Penape-Gunogu pair presents difficulty for our tested models, with systematic disagreement with humans by Gemini and GPT-4, suggesting the influence of additional information contained within the models' training data as to the connotations of these pseudowords.

4 Magnitude Symbolism

Whilst the Kiki-Bouba effect demonstrates sound symbolism in relation to the perceived spikiness/roundness of an object, magnitude symbolism refers to the non-arbitrary relationship between certain vowels and the perceived physical size of the entity they refer to and is commonly demonstrated through the names "Mil" and "Mal", where the high front vowel in "Mil" is associated with small entities, and the low back vowel of "Mal" is associated with larger entities.

[5]We see LLaVA variants demonstrate a clear preference for whatever pseudoword is presented in the first position.

Figure 4: Examples of "Mil"-style (tiny) and "Mal"-style (huge) generations with DALL-E 3. In total, 50 images were generated, with 25 per condition (the entities remaining constant). The ground truth is taken as the majority human vote.

4.1 Methodology

Image Dataset We follow the same process as §3.1, but use the characteristics of "tiny" and "huge" with the following prompt: "*Generate an image of a/an [ENTITY] in isolation, with something else to help judge scale/perspective*". We use the same noun entities as in §3.1. Example generations are in Figure 4.

Pseudowords As in the Shape Symbolism experiment (§3), we wish to mitigate potential bias from the memorisation capability of the VLMs. To this end, we use the "Mil" and "Mal" often associated with this test (Sapir, 1929) in addition to other phonetically similar pseudowords. Additionally, to avoid gross extraneous factors arising from using words that are meaningful in English (such as "mil" referring to millilitres, and "mal" being associated with badness, i.e., malpractice/malnourishment), we create the minimal pairs *Dil/Dal*, *Zil/Zal*, *Geel/Gaal*, *Beel/Baal*, *Weel/Waal*, and *Leel/Laal*. The former three exploit the contrast between /ɪ/ and /a/, whilst the latter exploit /i/ and /ɑ/, with a range of consonants for variation.

Task Setting We use the same setup as in §3.1 but present the models with one candidate from either Magnitude-based name category (i.e. "Mil"-esque or "Mal"-esque pseudowords), rather than Kiki/Bouba related pseudowords. We additionally provide an extended prompt we call *informed*, which prepends "*This task is related to the phenomenon of Magnitude Symbolism, which is a non-arbitrary relationship between the sound of a word and its association with size and scale*" to give additional task knowledge.

Models & Evaluation We use the same models and evaluation protocols as in §3.1.

4.2 Results

Overall, similar to §3 we see mixed results in Figure 5 as to which model performs best, but GPT-4 demonstrates higher levels of agreement across most conditions in a clearer pattern than what was seen in the Shape Symbolism experiment. In several cases, we see agreement that is significantly in line with human perception, with 90+% agreement (e.g., GPT-4 in the Zil-Zal and Weel-Waal conditions). When comparing the *standard* and *informed* prompts, we see a much more substantial increase in performance, indicating that the VLMs fundamentally understand the relationship between sound and perceived size, but were focused on other aspects of the provided imagery when not explicitly directed towards size in the *standard* condition. Regarding the LLaVA models, we see the mid-sized 13B variant outperforming the 7B and 34B models in most conditions (rather than performance increasing alongside parameter count).

5 Iconicity Ratings

In this section we investigate whether LLMs demonstrate human-like associations between word forms and the entities/concepts they symbolise.[6] Winter et al. (2023) present a dataset of 14k+ human judgements of iconicity to which we compare our LLM ratings (where, on a 7-point scale, "how" scored 1.3, whilst "woof" scored 6.8 due to being onomatopoeic).

5.1 Methodology

Models We use a range of modern LLMs for this task. Specifically, **GPT-4** (OpenAI et al., 2023), **GPT-3.5-Turbo** (Ouyang et al., 2022), **LLaMA-2** (7B, 13B and 70B) (Touvron et al., 2023), **FLAN-T5** (base and XL) (Raffel et al., 2023; Chung et al., 2022), and **Mistral-7B** (Jiang et al., 2023). Implementation details are presented in Appendix A.1.

[6] We treat this as VLMs "imitating" an understanding of sound symbolism, as they of course cannot actually hear.

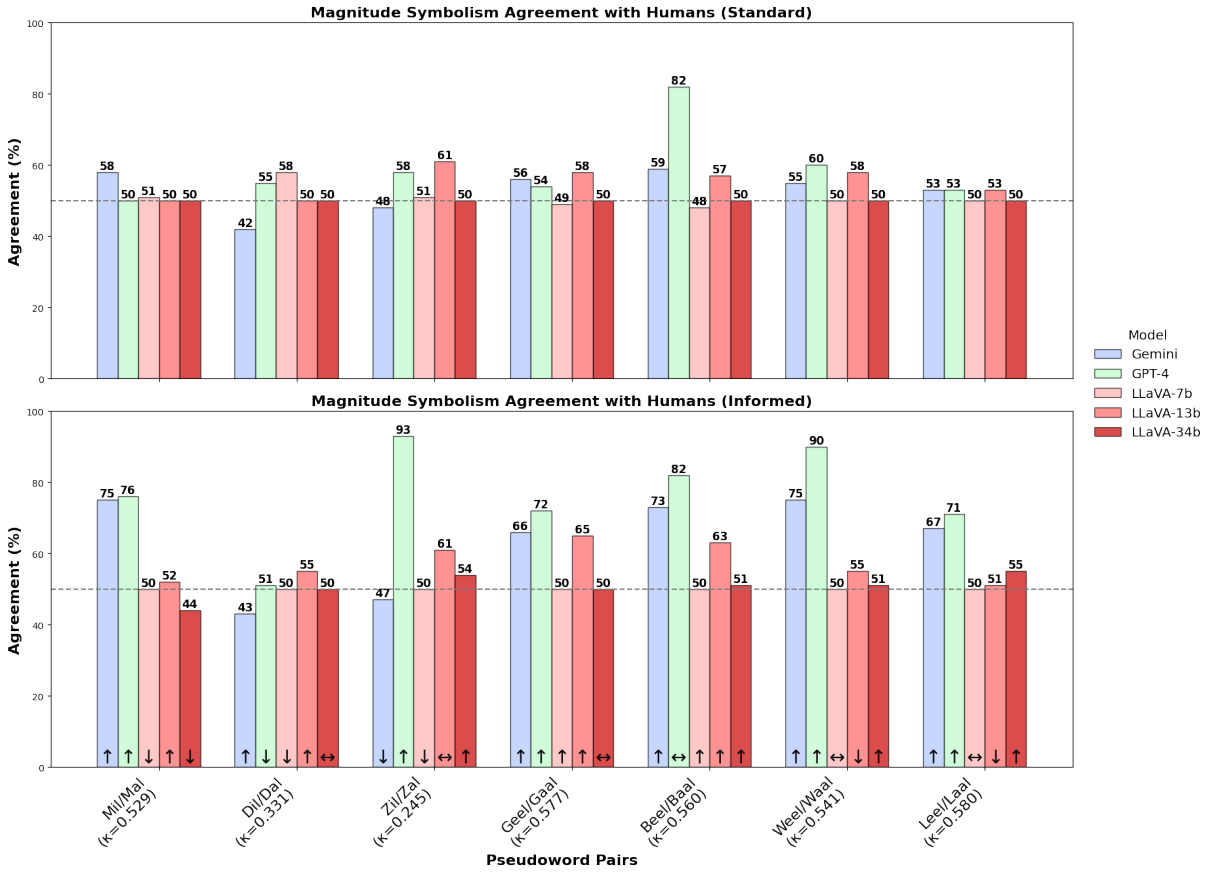

Figure 5: Results of the Magnitude Symbolism experiments per pseudoword pair. Fleiss' κ (Fleiss, 1971) for inter-annotator agreement between humans is presented next to each pseudoword pair. Arrows indicate the direction of agreement change from the standard prompt to the informed prompt. The dashed line represents 50%, akin to chance-level agreement. Full results in table form are presented in Table 3 within Appendix A.2. In all cases, we are comparing with the human majority vote.

Existing work by Trott (2024) has investigated whether or not GPT-4 is able to reflect human judgements and reported a Spearman correlation of 0.59. However, in trying to verify these findings, we see that our ratings differ quite strongly. For this reason, we re-run this experiment on the first 7k entries of Winter et al. (2023) and compare GPT-4 ratings as of late November 2023 with the human judgements in Winter et al. (2023) as well as the GPT-4 judgements from Trott (2024) that use an earlier version of GPT-4.[7] In the other cases, we use the entire 14,772 entry dataset, except for GPT-3.5-Turbo, where we remove 209 words that triggered OpenAI's safeguarding filters.

Prompting We adopt the same prompting strategy as Trott (2024). In doing so, we use a modified version of what was presented to human participants in Winter et al. (2023). In summary, we request ratings of iconicity on a 1-7 scale, where 1 is not at all iconic, and 7 is highly iconic. The full prompt presented to the models is available in Appendix C.2.

5.2 Results

Correlations between each LLM and human judgements are presented in Table 1. We observe that the ability of LLMs to rate the iconicity of English words appears to be dependent on model size. For instance, we see no true correlation with any of our FLAN T5 models (base = 250M, XL = 3B) or our smallest LLaMA-2 variant (7B), but we begin to see significant correlations with another 7B parameter model, that being instruction-tuned Mistral-7B. When investigating our larger models, we see the 13B LLaMA-2 variant demonstrate Spearman/Pearson correlations of .379 and .381, respectively. Interestingly, however, our largest LLaMA-2 model with 70B parameters performs worse at this task than the 13B variant as seen previously, with correlations of .304/.332. Regarding OpenAI models, GPT-3.5-Turbo demonstrates a

[7] We test only on the first 7k entries due to cost. Whilst GPT-4-Turbo is markedly cheaper, this would not be a true reproduction. Please note that GPT-4o was not released at the time.

moderate correlation with humans at .420/.439. Finally, GPT-4 presents the strongest correlations. However, we observe that (on the first 7k entries), the version of GPT-4 we use (late November 2023) performs worse than the earlier version presented by Trott (2024), further demonstrating how continuously updated models also require continuous evaluation due to receiving additional training data (Spearman correlations of 0.537 vs 0.575, respectively).

	Spearman		Pearson	
Model	Corr.	p	Corr.	p
FLAN-T5 Base	-.035	<.001	-.037	<.001
FLAN-T5 XL	.000	.991	-.002	.824
Mistral-7B-Instruct	.382	<.001	.377	<.001
LLaMA-2 7B	-.003	.687	.005	.544
LLaMA-2 13B	.379	<.001	.381	<.001
LLaMA-2 70B	.304	<.001	.332	<.001
GPT-3.5-Turbo	.420	<.001	.439	<.001
GPT-4 (Trott, 2024)	**.575**	<.001	**.615**	<.001
GPT-4 (Ours, Nov. '23)	**.537**	<.001	**.594**	<.001

Table 1: Correlations between human ratings from Winter et al. (2023) and LLMs. GPT-4 ratings were only collected for the first 7k examples due to cost. Trott (2024) report a Spearman correlation of .590 across the entire dataset. p refers to p-value.

6 Discussion

The Source of Sound Symbolism Across our experiments, we see evidence that LLM/VLMs are capable of making decisions that are similar to those of humans in sound symbolism tasks, whilst only having access to textual and visual modalities, while human decisions are believed to be grounded in sound. We hypothesise several reasons for the emergence of sound symbolism in LLM/VLMs.

Firstly, due to human languages exhibiting mostly regular orthography, auditory information in speech is moderately reflected in the spellings of words via grapheme sequences (a characteristic that grapheme-to-phoneme conversion models have long exploited). Through this, text-based models are able to learn associations between grapheme sequences and semantics, based on more abstract characteristics than morpheme combinations alone, such as phonaesthemes (Kaushal and Mahowald, 2022). Whilst such models have no embodied understanding of sound, such statistical patterns pose a viable signal for the implicit learning of sound-based phenomena.

Secondly, such associations between sounds (or grapheme combinations) and physical characteristics are naturally present in language, such as in poetry, narratives, or descriptions of entities that are cute, scary, small, or large, and are consequently paired with relevant visual stimuli in image captions when training vision modules for multimodal systems. However, such associations are subtle and not entirely ubiquitous. For example, whilst the high front vowel /ɪ/ typically associated with small entities is present in "tiny" and "mini", the word "small" itself possesses a low back vowel /ɔː/.

As a result, the relatively weak performance of our tested models could also be explained by the relative lack of sound-symbolism-heavy language in the models' training data which is overshadowed by more prosaic language forms that do not exploit these phenomena as readily. This in turn would explain why the closed-source models we tested (e.g. GPT-4/Gemini) outperform open-source models due to the significant (assumed) differences in parameter size, allowing the larger closed models to retain more information regarding sound symbolism within the weights, in addition to being continuously updated with RLHF.

The results of our multimodal experiments additionally demonstrate that under certain conditions, VLMs show systematic disagreement with human labels, indicating the potential interference of additional knowledge contained within language model training data that influences the associations made between pseudowords and images that are not present in humans. However, it is important to note that in our experiments we compare language model selections against the majority vote or mean scores assigned by humans. Consequently, this results in a comparison to an "ideal" human by necessity, overlooking individual differences in perception across humans (where for a decision to be "human-like", it has to match a choice made by any human, rather than the majority). Consequently, higher agreement levels can be observed when compared to the choices of individual humans, as inter-human agreement is not perfectly aligned in these tests.

Future Directions As a result of LLM/VLMs not being able to fully reflect human preferences in tasks regarding sound symbolism, it remains a promising future direction to explicitly pre-train language models on more sound-symbolism heavy datasets or explicitly include sound-symbolism-related tasks into the training or finetuning of these models for use on related downstream tasks (such as creative writing and marketing). Additionally, investigating the reason behind model predictions is a promising direction, such as through additional prompting to

generate textual justifications, or investigating the visual attention of VLMs to investigate whether they are attending to characteristics closely associated with the concepts being tested (e.g., spikes).

7 Conclusion

We have shed light on the processes underlying multimodal perception and understanding in language models. To do so, we performed a series of tests on modern VLM/LLMs regarding their ability to exhibit an understanding of sound symbolism. Through comparison with human judgements, we see that VLMs are able to approximate human perception in sound symbolism tests under certain conditions, such as when informed of the nature of the study (via the *informed* prompts), but struggle overall. We additionally see that magnitude symbolism potentially presents an easier pattern for VLMs to recognise than shape symbolism, with selections having a higher agreement with humans on Magnitude Symbolism tests than Shape Symbolism. We also see that the ability of LLMs to emulate human judgements of iconicity scales more linearly with model size. These findings indicate room for future research on more explicit inclusion of abstract perceptual properties into language model training in order to facilitate better *in silico* experimentation and improve performance on other downstream tasks.

Limitations

Owing to the relatively small sample sizes (i.e., the number of pseudoword pairs in the VLM-related tasks), we treat this work as a proof-of-concept as to the ability of LLMs to perform well in the tasks we present and encourage other parties to engage in similar research at scale if their situation permits. Additionally, whilst sound symbolism is believed to be largely language agnostic, we only use native English speakers and pseudowords that are phonotactically legal in English in the present work. Additionally, some of our chosen pseudowords are taken from existing literature. Whilst we investigated the prevalence of these words in the context of sound symbolism within internet resources in order to mitigate memorisation from training data, it remains possible that some level of data contamination may be present (although the overall low performance casts doubt on this). Furthermore, we present only the orthographic forms of the pseudowords to human participants, resulting in potential variation between speakers regarding phonetic realisation.

Ethics Statement

We believe in and firmly adhere to the Code of Conduct in the performance of this work and the methods involved. All of our imagery generations were provided via accessing the respective OpenAI APIs, and in discovering imagery that triggered OpenAI's built-in guardrails, we replaced these images with other options. All human evaluation was performed by consenting adult participants who were provided with a participant information sheet and subsequently signed a consent form in line with the Ethics procedures of the primary author's institution (who approved the ethical validity of the study performed herein). Additionally, we present this work as a demonstration of interesting behaviours within (very) large LLMs, but do not condone the wholesale replacement of human participants in related psycholinguistic/cognitive/psychological experimentation, but rather view *in silico* experimentation as a useful tool primarily for prototyping.

Acknowledgments

Tyler Loakman is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

Visual elements in Figure 1 were taken from Flaticon. Specifically, banners (SANB), brain (Freepik), GPU (Taufik Ramadhan), "versus" (Afif Fudin), descending/ascending (Mie Nakae), and thought bubble (Aranagraphics).

References

Ekaterina Abramova and Raquel Fernández. 2016. Questioning arbitrariness in language: a data-driven study of conventional iconicity. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 343–352, San Diego, California. Association for Computational Linguistics.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal,

Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image geneartion with better captions.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *ArXiv*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, Gary Lupyan, Grace E. Oh, Jing Paul, Caterina Petrone, Rachid Ridouane, Sabine Reiter, Nathalie Schümchen, Ádám Szalontai, Özlem Ünal-Logacev, Jochen Zeller, Marcus Perlman, and Bodo Winter. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841). Publisher Copyright: © 2021 The Authors.

Ferdinand de Saussure and Wade Baskin. 2011. *Course in General Linguistics: Translated by Wade Baskin. Edited by Perry Meisel and Haun Saussy*. Columbia University Press.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

JL Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Shailee Jain, Vy A. Vo, Leila Wehbe, and Alexander G. Huth. 2023. Computational Language Modeling and the Promise of in Silico Experimentation. *Neurobiology of Language*, pages 1–27.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7b. *ArXiv*.

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

Seth Ketron and Nancy Spears. 2021. Sound-symbolic signaling of online retailer sizes: The moderating effect of shopping goals. *Journal of Retailing and Consumer Services*, 58:102245.

W. Köhler. 1929. *Gestalt psychology.* Gestalt psychology. Liveright, Oxford, England.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Nelson F. Liu, Gina-Anne Levow, and Noah A. Smith. 2018. Discovering phonesthemes with sparse regularization. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 49–54, New Orleans. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2023. TwistList: Resources and baselines for tongue twister generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–589, Toronto, Canada. Association for Computational Linguistics.

Tyler Loakman, Chen Tang, and Chenghua Lin. 2024. Train & constrain: Phonologically informed tongue-twister generation from topics and paraphrases. *arXiV*.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(3).

Kosuke Motoki, Toshiki Saito, Jaewoo Park, Carlos Velasco, Charles Spence, and Motoaki Sugiura. 2020. Tasting names: Systematic investigations of taste-speech sounds associations. 80.

Alan K. S. Nielsen and Drew Rendall. 2013. Parsing the role of consonants versus vowels in the classic takete-maluma phenomenon. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 67(2):153–163.

Valeria Occelli, Gianluca Esposito, Paola Venuti, Giuseppe Maurizio Arduino, and Massimiliano Zampini. 2013. The takete—maluma phenomenon in autism spectrum disorders. *Perception*, 42(2):233–241. PMID: 23700961.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ananya Passi and S. P. Arun. 2022. The bouba–kiki effect is predicted by sound properties but not speech properties. *Attention, Perception, & Psychophysics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

V S Ramachandran and Edward M Hubbard. 2001. Synaesthesia — a window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*.

Valentino Sabbatino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland. Association for Computational Linguistics.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225–239.

David M. Sidhu, Angeliki Athanasopoulou, Stephanie L. Archer, Natalia Czarnecki, Suzanne Curtin, and Penny M. Pexman. 2023. The maluma/takete effect is late: No longitudinal evidence for shape sound symbolism in the first year. *PLOS ONE*, 18(11):1–23.

David M. Sidhu and Gabriella Vigliocco. 2023. I don't see what you're saying: The maluma/takete effect does not depend on the visual appearance of phonemes as they are articulated. *Psychonomic Bulletin & Review*, 30(4):1521–1529.

David M. Sidhu, Chris Westbury, Geoff Hollis, and Penny M. Pexman. 2021. Sound symbolism shapes the english language: The maluma/takete effect in english nouns. *Psychonomic Bulletin & Review*, 28(4):1390–1398.

Charles Spence. 2012. Managing sensory expectations concerning products and brands: Capitalizing on the potential of sound and shape symbolism. *Journal of consumer psychology*, 22(1):37–54.

Suzy J. Styles and Lauren Gawne. 2017. When does maluma/takete fail? two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception*, 8(4). PMID: 28890777.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*.

Dmitri Usnadze. 1924. Ein experimenteller beitrag zum problem der psychologischen grundlagen der namengebung. pages 24–43.

Bodo Winter, Gary Lupyan, Lynn K. Perry, Mark Dingemanse, and Marcus Perlman. 2023. Iconicity ratings for 14,000+ english words. *Behavior Research Methods*.

Ivan P. Yamshchikov, Viascheslav Shibaev, and Alexey Tikhonov. 2019. Dyr bul shchyl. proxying sound symbolism with word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 90–94, Minneapolis, USA. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

Kiki-Bouba & Mil-Mal For the Kiki-Bouba and Mil-Mal sound symbolism experiments, we access *gemini-pro-vision* via the Google Gemini API. For GPT-4, we use *gpt-4-vision-preview* via the OpenAI Chat Completions API. For our open-source LLaVA models at various sizes, we specifically use *llava-v1.6-vicuna-7b-hf*, *LLaVA-v1.6-vicuna-13b-hf* and *LLaVA-v1.6-vicuna-34b-hf* from publicly available checkpoints on Hugging Face. We use default hyperparameters for all models to test their "off-the-shelf" capability. Human participants were shown generated imagery with the binary pseudo-word options via Google Forms. The order of image presentation was randomised, whilst the order of the pseudowords was kept static. Separate Google Forms were used for each pseudoword pair. Participants were able to complete the forms at their own pace within a period of approximately 2 weeks.

Iconicity Ratings For our iconicity rating experiments, we use the following models from Hugging Face: FLAN-T5 Base (*google/flan-t5-base*), FLAN-T5 XL (*google/flan-t5-xl*) Mistral-7B (*mistralai/Mistral-7B-Instruct-v0.2*), LLaMA-2 (*meta-llama/Llama-2-7b-hf, meta-llama/Llama-2-13b-hf, meta-llama/Llama-2-70b-hf*). We access GPT-3.5-Turbo and GPT-4 via the OpenAI Chat Completions API. We also keep all model hyperparameters at default settings. For the LLaVA models, we modify the prompt slightly by adding choice labels (A/B) rather than requesting the pseudoword itself to be returned in order to directly access single-token output probabilities.

Entities Within our DALL-E 3 generations in the aforementioned experiments, we select the following list of entities in order to have a range of characteristics (including animate and inanimate entities): *alien, bed,*

Figure 6: Examples of ImageNet and DALL-E 3 image pairings for both the soundscape description and backtranslation experiments. Here, the GPT-4 description of the ImageNet image has been used in a prompt to generate the novel DALL-E 3 image.

bird, bottle, cat, chair, desk, dog, door, fish, flower, fruit, ghost, house, insect, lizard, machine, person, plane, planet, snake, toy, tree, vegetable, and *vehicle*.

A.2 Full Results

Shape Symbolism (Kiki/Bouba) Table 2 presents the full unabridged results for the Sound Symbolism experiments presented in §3.

Magnitude Symbolism (Mil/Mal) Table 3 presents the full unabridged results for the Magnitude Symbolism experiments presented in §4.

B Soundscape Description

We additionally ask to what extent a VLM (specifically GPT-4) is able to demonstrate a sense of "hearing" via being tasked with describing a perceived soundscape (including the use of onomatopoeia) from an image. We use images from 2 different sources for the following experiments.

B.1 Methodology

Real-World Firstly, due to requiring high-quality publicly available images to represent the real-life condition, we utilise a subset of ImageNet.[8] To select our candidate images, 2 authors of this work selected a list of images that are believed to represent a wide range of soundscapes (e.g. a peaceful beach, violent waves, a desert, a car, a plane, etc.), These 50 were selected from a unique set of images that all represented different classes under the ImageNet taxonomy. Importantly, some of the chosen images were not usable with GPT-4 due to containing entities that trigger OpenAI's safeguarding restrictions (such as an image of a baby in a cot, or a couple of hunters with rifles). In such cases, we replace these images with alternative selections that the 2 authors agree are high quality. [9]

Generative AI For our GenAI-based imagery, we use scene descriptions from GPT-4 (which are generated as part of the output for this task) of the ImageNet imagery and prompt DALL-E 3 to generate images from these descriptions via the OpenAI API. In doing so, we then create a parallel dataset of 25 real-world images, 25 LLM descriptions of said images, 25 DALL-E 3 generations using the aforementioned LLM descriptions, and finally, 25 LLM descriptions of the DALL-E 3 generations. This therefore allows us to ensure that our testing is robust to novel images, as ImageNet imagery is likely to have been a part of the training set for GPT-4's vision module. Additionally, this also facilitates an investigation as to how consistent GPT-4 is at scene description and generation (analogous to testing Neural Machine Translation via back-translation).

Task Setting We perform this task in the following way. For each condition (ImageNet/DALL-E), we present GPT-4 with the following prompt: "*Imagine that the provided image is a window to another world. Describe the scene in 3 paragraphs discussing the following aspects: Paragraph 1: Describe what you see in the image, including the entities and the perceived environment. Paragraph 2: Describe what you hear in the image (i.e. the soundscape), including sounds from the identified entities, as well as the perceived environment. Paragraph 3: In reference to the sounds mentioned in Paragraph 2, describe these sounds using onomatopoeia (i.e. words that sound like the sounds you are trying to describe). Provide your answer to Paragraph 3 as a series of bulletpoints.*".

[8] Specifically *imagenet-1k*, available at https://huggingface.co/datasets/imagenet-1k

[9] For example, some ImageNet images are hard to decipher as they are intended to test the capabilities of computer vision models. We opt to avoid such instances and show preference towards images that present a full scene, as opposed to a single object.

Prompt	Kiki/Bouba (κ .731)				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	60% (κ .200)	**86% (κ .720)**	50% (κ .000)	44% (κ -.120)	50% (κ .000)
Informed	63% (κ .260)	**88% (κ .760)**	50% (κ .000)	46% (κ -.080)	53% (κ .060)
Prompt	**Kalika/Mabobe (κ .716)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	43% (κ -.140)	**58% (κ .160)**	50% (κ .000)	52% (κ .040)	50% (κ .000)
Informed	63% (κ .260)	**66% (κ .320)**	50% (κ .000)	55% (κ .100)	51% (κ .020)
Prompt	**Zaki/Umbu (κ .753)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b ↓**	**LLaVA-34b ↑**
Standard	**64% (κ .277)**	58% (κ .152)	50% (κ .002)	54% (κ .103)	50% (κ .000)
Informed	**73% (κ .464)**	70% (κ .402)	48% (κ .000)	52% (κ .065)	51% (κ .019)
Prompt	**Tiki/Giba (κ .655)**				
	Gemini ↓	**GPT-4 ↓**	**LLaVA-7b**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	49% (κ -.020)	43% (κ -.140)	50% (κ .000)	**58% (κ .160)**	50% (κ .000)
Informed	50% (κ .000)	36% (κ -.280)	50% (κ .000)	**63% (κ .260)**	51% (κ .020)
Prompt	**Kitaki/Gugagu (κ .670)**				
	Gemini ↑	**GPT-4 ↓**	**LLaVA-7b ↑**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	41% (κ -.180)	47% (κ -.060)	48% (κ -.040)	47% (κ -.060)	**50% (κ .000)**
Informed	54% (κ .080)	43% (κ -.140)	51% (κ .020)	50% (κ .000)	**57% (κ .140)**
Prompt	**Hatiha/Bodubo (κ .706)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	**61% (κ .220)**	47% (κ -.060)	50% (κ .000)	46% (κ -.080)	50% (κ .000)
Informed	**78% (κ .560)**	52% (κ .040)	50% (κ .000)	54% (κ .080)	51% (κ .020)
Prompt	**Penape/Gunogu (κ .627)**				
	Gemini ↓	**GPT-4 ↓**	**LLaVA-7b ↑**	**LLaVA-13b ↓**	**LLaVA-34b**
Standard	38% (κ -.240)	29% (κ -.420)	47% (κ -.060)	38% (κ -.240)	**50% (κ .000)**
Informed	26% (κ -.480)	22% (κ -.560)	50% (κ .000)	35% (κ -.300)	**50% (κ .000)**
Prompt	**ALL (excl. Kiki/Bouba)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	49.33% (κ -.014)	47.83% (κ -.045)	49.17% (κ -.016)	49.71% (κ -.013)	**50.40% (κ .012)**
Informed	**57.33% (κ .147)**	48.17% (κ -.036)	50.17% (κ .003)	51.50% (κ .034)	51.83% (κ .037)

Table 2: Results of the Shape Symbolism experiments per pseudoword pair. Fleiss' κ (Fleiss, 1971) for inter-annotator agreement between humans is presented next to each pseudoword pair. For each VLM and word pair, we present Cohen's κ for agreement between the models and the human majority vote (Cohen, 1960). The model with the highest agreement per prompt is in **bold**, and the best performing open-source model (i.e., LLaVA variant) is underlined. Arrows next to model names indicate the direction of agreement change from the standard prompt to the informed prompt.

Following this, we ask 5 human evaluators (a subset from the main experiments), to evaluate the 3 paragraphs on a 1-5 scale, where 1 = very bad, and 5 = excellent (i.e., one rating for the visual description, one for the soundscape description, and one for the assignment of onomatopoeia to the soundscape). The instructions presented to human participants for this task are presented in Appendix C.3.[10] The order of image presentation to participants is randomised to avoid order effects.

[10]We present detailed instructions in order to moderate the understanding of what we would consider the different ratings to be indicative of in order to minimise individual perceptions of the instructions.

B.2 Results

The results of the soundscape description task can be seen in Table 4 and an example generation is presented in Figure 7. Overall, it can be seen that human evaluators thought positively of all 3 elements asked for from GPT-4, including the visual description (which would explain performance in the following section), soundscape description and onomatopoeia, with all criteria averaging at least 4. This therefore demonstrates that GPT-4 is able to provide convincing descriptions of auditory experiences when provided with a valid image. One key thing to note is that whilst the standard deviations are consistently low, onomatopoeia demonstrates the lowest consistently. This is to be ex-

Prompt	Mil/Mal (κ .529)				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b ↑**	**LLaVA-34b ↓**
Standard	**58% (κ .152)**	50% (κ .031)	51% (κ .018)	50% (κ .000)	50% (κ .000)
Informed	75% (κ .512)	**76% (κ .529)**	50% (κ .000)	52% (κ .043)	44% (κ -.109)
Prompt	**Dil/Dal (κ .331)**				
	Gemini ↓	**GPT-4 ↓**	**LLaVA-7b ↓**	**LLaVA-13b ↑**	**LLaVA-34b**
Standard	42% (κ -.168)	55% (κ .098)	**58% (κ .149)**	50% (κ .000)	50% (κ .000)
Informed	43% (κ -.128)	51% (κ .010)	50% (κ .000)	**55% (κ .122)**	50% (κ .000)
Prompt	**Zil/Zal (κ .245)**				
	Gemini ↓	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b**	**LLaVA-34b ↑**
Standard	48% (κ -.003)	58% (κ .149)	51% (κ .022)	**61% (κ .222)**	50% (κ .000)
Informed	47% (κ -.031)	**93% (κ .860)**	50% (κ .000)	61% (κ .207)	54% (κ .083)
Prompt	**Geel/Gaal (κ .577)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b ↑**	**LLaVA-34b**
Standard	56% (κ .120)	54% (κ .080)	49% (κ -.020)	**58% (κ .160)**	50% (κ .000)
Informed	66% (κ .320)	**72% (κ .440)**	50% (κ .000)	65% (κ .300)	50% (κ .000)
Prompt	**Beel/Baal (κ .560)**				
	Gemini ↑	**GPT-4**	**LLaVA-7b**	**LLaVA-13b ↓**	**LLaVA-34b**
Standard	59% (κ .180)	**82% (κ .640)**	48% (κ -.040)	57% (κ .140)	50% (κ .000)
Informed	73% (κ .460)	**82% (κ .640)**	50% (κ .000)	63% (κ .260)	51% (κ .020)
Prompt	**Weel/Waal (κ .541)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b**	**LLaVA-13b ↓**	**LLaVA-34b ↑**
Standard	55% (κ .100)	**60% (κ .200)**	50% (κ .000)	58% (κ .160)	50% (κ .000)
Informed	75% (κ .500)	**90% (κ .800)**	50% (κ .000)	55% (κ .100)	51% (κ .020)
Prompt	**Leel/Laal (κ .580)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b**	**LLaVA-13b ↓**	**LLaVA-34b ↑**
Standard	**53% (κ .060)**	**53% (κ .060)**	50% (κ .000)	**53% (κ .060)**	50% (κ .000)
Informed	67% (κ .340)	**71% (κ .420)**	50% (κ .000)	51% (κ .020)	55% (κ .100)
Prompt	**ALL (excl. Mil-Mal)**				
	Gemini ↑	**GPT-4 ↑**	**LLaVA-7b ↓**	**LLaVA-13b ↑**	**LLaVA-34b ↑**
Standard	52.17% (κ .048)	**58.50% (κ .169)**	51.00% (κ .019)	56.17% (κ .127)	50.00% (κ .000)
Informed	61.83% (κ .243)	**76.50% (κ .528)**	50.00% (κ .000)	58.33% (κ .168)	51.83% (κ .037)

Table 3: Results of the Magnitude Symbolism experiments per pseudoword pair. Fleiss' κ (Fleiss, 1971) for inter-annotator agreement between humans is presented next to each pseudoword pair. For each VLM, we present Cohen's κ for agreement between the models and the human majority vote (Cohen, 1960). The model with the highest agreement per prompt is in **bold**, and the best performing open-source model (i.e., LLaVA variant) is underlined. Arrows next to model names indicate the direction of agreement change from the standard prompt to the informed prompt.

pected when evaluating a literary device, as different people may have different preferences regarding onomatopoeia they would use for certain circumstances. Additionally, there may be cases where GPT-4 has described something such as a stream and assigned the onomatopoeia "whoosh", which to one individual may sound too aggressive and resemble fast-moving water, when their own interpretation of a stream is more gentle (perhaps better suiting "lap lap").

C GPT-4 Image "Backtranslation"

We use the images and descriptions we collected to test the internal consistency of the OpenAI pipeline.[11]

Soundscape Descriptions		
	Mean	**SD**
IN Visual	4.19	0.46
IN Soundscape	4.42	0.33
IN Onomatopoeia	4.24	0.47
D3 Visual	4.59	0.39
D3 Soundscape	4.55	0.36
D3 Onomatopoeia	4.24	0.46

Table 4: Average ratings given to the visual, soundscape, and onomatopoeia descriptions given by GPT-4 across 2 conditions. *IN* refers to images from ImageNet, and *D3* refers to DALL-E 3 generations.

[11] As of late November 2023.

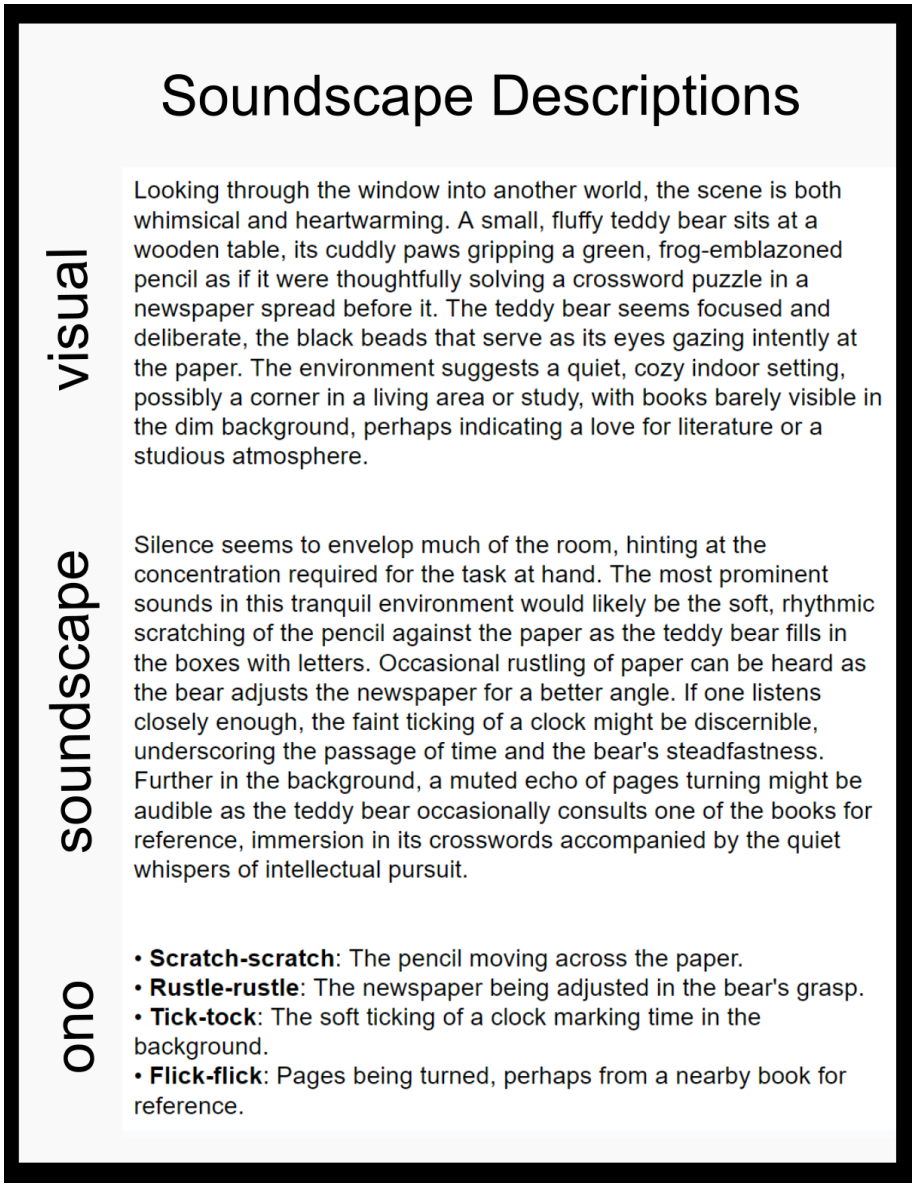

Figure 7: An example output from GPT-4 when asked to describe the visuals, soundscape, and perceived onomatopoeia of an image. In this case, the image is the ImageNet generation of a teddy solving a crossword seen in Figure 6. *ono* refers to onomatopoeia.

In effect, our newly generated images demonstrate a process analogous to the back-translation used in Neural Machine Translation. To test the consistency of this process, we ask our evaluators to rate the generations on the following criteria: To what extent does the DALL-E 3 generated image present the same visual scene as the original ImageNet source image (on a scale of 1 to 5, where 1 = barely related, and 5 = all the main elements are captured).[12]

[12]We specify to evaluators that this task is indifferent to changes in art style, as we have not specified to DALL-E 3 that its generations should be photorealistic.

C.1 Results

Table 5 presents the results of our human evaluation. As we can see, the consistency of the pipeline is viewed favourably, with a mean rating of 4.18 across the 25 image pairings and a low standard deviation of 0.49. Importantly, no comparison was rated lower than 3 by any evaluator. This result is quite surprising given the $\tilde{1}$00-word descriptions provided by GPT-4, indicating that GPT-4-vision is highly capable of noticing the most salient aspects of any image for recreation. The result of automatic evaluation comparing the descriptions from ImageNet and DALL-E 3 images are presented in Table 6, echoing a similar pattern to the human evaluation.

Ratings	
Mean	**SD**
4.18	0.49

Table 5: Average human ratings given to the consistency of the generation pipeline when using GPT-4 descriptions of an ImageNet image to prompt DALL-E 3 to replicate. We refer to DALL-E 3 as D3 and ImageNet as IN.

Automatic Evaluation				
BS-P	**BS-R**	**BS-F**	**R-L**	**BLEU**
0.84	0.80	0.82	0.13	0.04

Table 6: Comparison across the visual, soundscape, and onomatopoeia descriptions from GPT-4 with the ImageNet condition as the reference and the DALL-E 3 condition as the prediction. BS-P/R/F stands for BERTScore Precision/Recall/F1, respectively. R-L is ROUGE Longest Common Subsequence. Hugging Face implementations were used for all metrics.

C.2 Full Iconicity Prompt

The full prompt provided to LLMs in the iconicity rating experiment is presented in Figure 8.

C.3 Materials Provided to Participants

Figure 9 presents the instructions given to participants when rating the consistency of the OpenAI GPT-4/DALL-E 3 pipeline, whilst Figure 10 presents the instructions presented to participants in the soundscape rating experiment. For Kiki-Bouba and Mil-Mal, the setup was straightforward, and participants were simply asked to select the name they believed to be the most appropriate.

“Some English words sound like what they mean. These words are iconic. You might be able to guess the meaning of such a word even if you did not know English.

Some words that people have rated high in iconicity are “screech,” “twirl,” and “ooze” because they sound very much like what they mean.

Some words that people have rated moderate in iconicity are “porcupine,” “glowing,” and “steep,” because they sound somewhat like what they mean.

Some words rated low in iconicity are “menu,” “amateur,” and “are,” because they do not sound at all like what they mean.

In this task, you are going to rate words for how iconic they are. You will rate each word on a scale from 1 to 7. A rating of 1 indicates that the word is not at all iconic and does not at all sound like what it means. 7 indicates that the word is high in iconicity and sounds very much like what it means.

It is important that you say the word out loud to yourself, and that you think about its meaning.

If you are unsure of the meaning or the pronunciation of a word, you have the option of skipping it.

Try to focus on the word meaning of the whole word, rather than decomposing it into parts. For example, when rating ‘butterfly’ think of the insect rather than “butter” and “fly,” and rate how well the whole meaning relates to the sound of the whole word “butterfly.”

On a scale from 1 (not iconic at all) to 7 (very iconic), how iconic is the word '{word}'?

Rating: "

Figure 8: The prompt provided to our LLMs in the iconicity judgement experiment.

Pipeline Consistency Experiment

For each question, please compare the real-world image to the generated image and **rate the image faithfulness on a 1-5 scale** (**1 = not at all faithful**, i.e. these images are unrelated, 5 = very faithful, these images present the same entities in the same environment doing the same activity with no major deviations)

- **Faithfulness** - Whether or not the image depicts the same essential entities, environment and atmosphere.

For this task do not penalise images for differences in art-style, as the generations are not intended to be photorealistic.

Additionally, do not treat the task as "spot the difference" over minor elements. Effectively, you are evaluating whether or not, if you were to describe the real-world image to an artist, and the generation is what they produced, you believe the generation matched the specification you had provided based on what you believe the key elements of the original image were.

"Examples"

- **Strong 5/5** - All main elements (entities, environment and atmosphere) are the same as the reference image. You would accept this as an artist's replication of an existing image with very minor changes (ignoring art style entirely)
- **Weak 5/5** All main elements (entities, environment and atmosphere) are the same as the reference image. The art style might be different, and things are not identical (slight outfit change, such as smart casual vs athletic wear, or an ambulance vs a police car, but not in a way that distracts from the essential information).
- **4/5 -** Some minor elements differ. For example, a bear is catching a fish in the reference, and in the generation the bear is stood alone (so it is no longer depicting the activity of hunting), but the environment/atmosphere and the main entity (the bear) is the same.
- **3/5** - The scenes are similar, but more major elements are different. For example, one image may be of a tropical beach in summer, whilst the other has a tropical beach in a thunderstorm. If you view the sunny weather in the reference as an essential element, you may criticise this strongly.
- **2/5** - Images that have at least one thing in common with the reference. For example, one image may be of a baby in a cot, and another has a toddler playing with toys. Here the main entity (the child) is consistent, but the atmosphere and environment differ heavily.
- **1/5** - The images have nothing reasonably in common, so that you would not assume a link. For example, if the reference is a field of cows, and the generation is a busy motorway with trucks.

Spend approximately 5-10 seconds on each.

Figure 9: Instructions provided to participants when asking to rate the consistency between ImageNet and DALL-E 3.

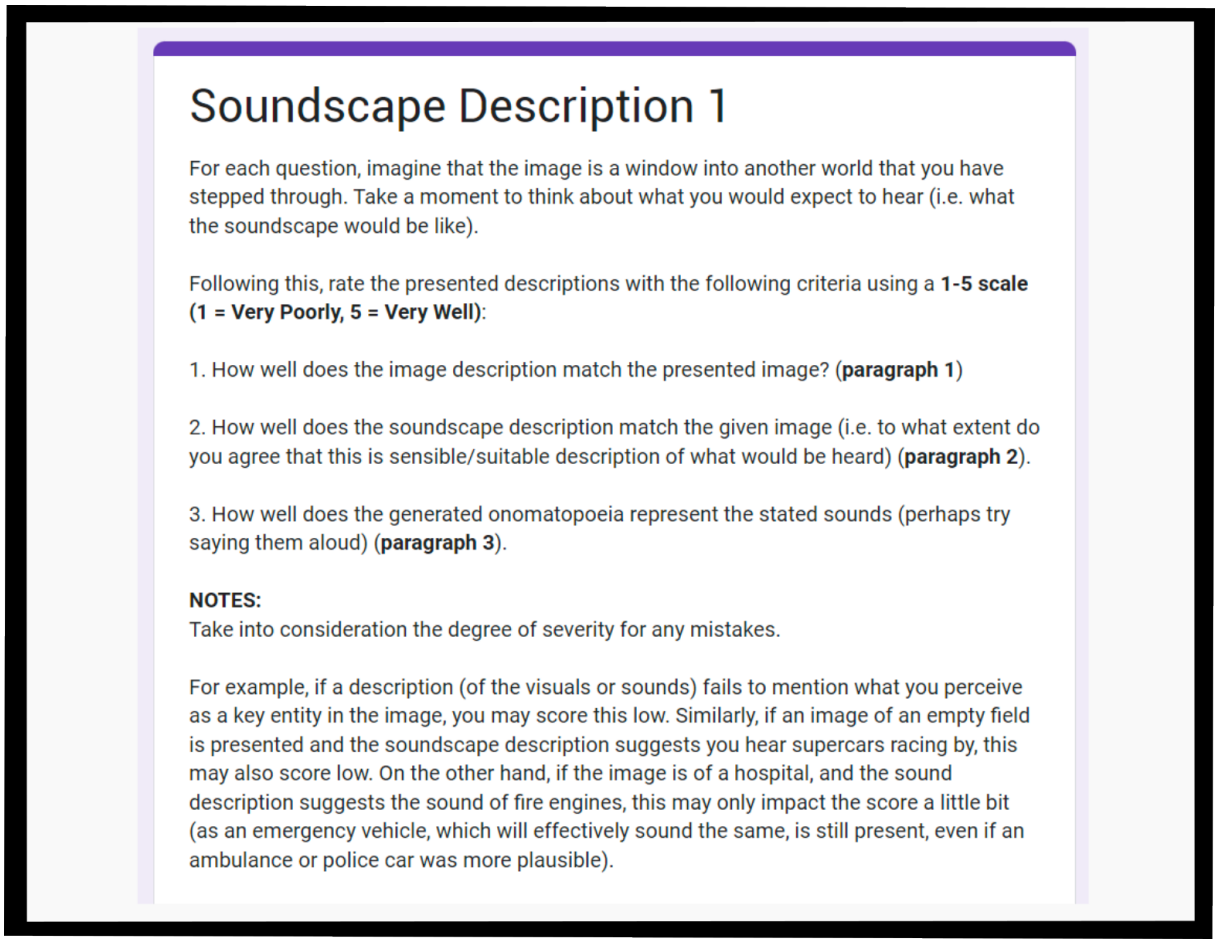

Figure 10: Instructions provided to participants when asking to rate the quality of the descriptions provided by GPT-4.