Multimodal Language Models Show Evidence of Embodied Simulation

Anonymous submission

Abstract

Multimodal large language models (MLLMs) are gaining popularity as partial solutions to the “symbol grounding problem” faced by language models trained on text alone. However, little is known about whether and how these multiple modalities are *integrated*. We draw inspiration from analogous work in human psycholinguistics on *embodied simulation*, i.e., the hypothesis that language comprehension is *grounded* in sensorimotor representations. We show that MLLMs are sensitive to *implicit* visual features like object shape (e.g., “The egg was in the skillet” implies a frying egg rather than one in a shell). This suggests that MLLMs activate implicit information about object shape when it is implied by a verbal description of an event. We find mixed results for color and orientation, and rule out the possibility that this is due to models’ insensitivity to those features in our dataset overall. We suggest that both human psycholinguistics and computational models of language could benefit from cross-pollination, e.g., with the potential to establish whether grounded representations play a *functional* role in language processing.

Keywords: grounding, multimodal language models, embodiment

1. Introduction

Recent advances in Large Language Models (LLMs) have generated an explosion of interest in their underlying capabilities and limitations (Thirunavukarasu et al., 2023). One oft-cited limitation of contemporary LLMs is that they are trained on linguistic input alone (Bender and Koller, 2020), and thus, unlike humans, lack access to embodied experience—seen by some as a prerequisite for language understanding (Bisk et al., 2020; Harnad, 1990; Mollo and Millière, 2023). Multimodal Large Language Models (MLLMs Driess et al., 2023; Girdhar et al., 2023; Huang et al., 2023)—which learn to associate linguistic representations with data from other modalities—may be a partial solution to this *symbol grounding problem* (Harnad, 1990). Yet despite impressive performance by MLLMs (Dosovitskiy et al., 2021), little is known about how distinct modalities (e.g., language and vision) are *integrated* within a model’s representational space, as they appear to be in humans.

We address this gap by turning to an analogous debate about the extent to which *human* semantic representations are grounded in sensorimotor experience (Barsalou, 1999). The *embodied simulation hypothesis* (Bergen, 2015; Glenberg, 2010) argues that language understanding involves the activation of grounded representations, i.e. that the same neural tissue recruited to perceive or participate in an event (e.g., kicking a soccer ball) is also engaged to understand language about that event (e.g., “She kicked the ball”). Indeed, a wide body of experimental evidence suggests that some degree of sensorimotor activation occurs during language processing (Zwaan and Pecher, 2012; Winter and Bergen, 2012). While there is ongoing debate about the functional relevance of embodied simulation (Glenberg et al., 2008; Mahon and Caramazza, 2008; Montero-Melis et al., 2022; Ostarek and Bottini, 2021), the evidence points to some degree of *cross-talk* between linguistic and sensorimotor neural systems.

Much of this evidence comes from the *sentence-picture verification task* paradigm (Stanfield and Zwaan, 2001). In this task, participants read a short sentence (e.g., “He hammered the nail into the wall”), then see a picture of an object (e.g., a nail) and must decide whether the object was mentioned in the preceding sentence. Crucially, when the image of the object *matches* the orientation (or shape, color, etc.) implied by the sentence (e.g., the nail is horizontal rather than vertical), participants are faster and more accurate in their decisions (Stanfield and Zwaan, 2001; Pecher et al., 2009; Connell, 2007). Because the object is the same (e.g., an egg), humans must be inferring visual features based on properties of the event itself (e.g., an egg cooking in a skillet).

In the current work, we applied these methodological insights to improve our understanding of MLLMs. We ask whether MLLM’s internal representations of linguistic input (e.g., "He hammered the nail into the wall") are more similar to representations of images that *match* visual features implied by that input than those that do not. To address this question, we adapted materials from three psycholinguistic studies that provide evidence for simulation of the implied orientation (Stanfield and Zwaan, 2001), shape (Pecher et al., 2009), and color (Connell, 2007) of objects. Note that this approach differs from a standard classification task: rather than classifying images on the basis of which objects they contain (e.g., “a cup of coffee”) or explicit features of those objects (e.g., “a black cup of coffee”), we are asking whether the MLLM activates *implicit* features that could be inferred from a more holistic event representation (e.g., “Joanne never took milk in her coffee” implies that the coffee is black).

2. Methods

2.1. Materials

We used stimuli from three experiments that measured visual simulation in human participants. Items were organized as quadruplets, consisting of a pair of images and a pair of sentences. Sentence pairs differed by implying that an object had a certain visual property (SHAPE, COLOR, or ORIENTATION). Each of the images in a pair matched the implied visual feature in one of the sentences (and therefore mismatched the other, see Figure 1).

60 quadruplets from Pecher et al. (2009) varied the implied SHAPE of an object. A sentence such as “There was an egg in the [refrigerator/skillet]” implied that the egg was either in its shell or cracked open. A pair of black-and-white images of eggs matched one of these sentences by displaying the relevant visual feature. Connell (2007) collected 12 quadruplets that vary the implied COLOR of an object. “Joanne [never/always] took milk in her coffee” implies black/brown coffee. The images differed only in color. Finally, Stanfield and Zwaan (2001) collected 24 quadruplets of sentences implying different ORIENTATIONS of an item, and line-drawings that were rotated to match the implied orientation. For instance “Derek swung his bat as the ball approached” suggests a horizontal bat, while “Derek held his bat high as the ball approached” suggests a vertical bat.

2.2. Model Evaluation

To probe MLLMs, we implemented a computational analogue of the sentence-picture verification task. Our primary question was whether a model’s representation of a given linguistic input (e.g., "He hammered the nail into the wall") was more similar to its representation of an image that matched an implied visual feature (e.g. horizontal orientation) compared to an image that did not (e.g. a vertical nail). For each sentence-image pair, we found the dot product between the MLLM embedding of the sentence and the image. This value quantifies the similarity between the linguistic and visual representations within the model. The dot product values were then passed through a softmax function, converting them into probabilities of the model associating eachimage with a given sentence:

$$p_{ij} = \frac{\exp(S_i \cdot I_j)}{\sum_{k=1}^{2} \exp(S_i \cdot I_k)}$$

where S_i is the embedding for sentence i, I_j is the embedding for image j, and p_{ij} is the softmax probability that sentence i matches with image j. To statistically evaluate the model’s performance, we conducted a t-test to compare the probabilities of matching (e.g., p_{11} and p_{22}) against mismatching (e.g., p_{12} and p_{21}) sentence-image pairs. A significant result, where the matching probabilities are greater than mismatching ones, would indicate that the MLLM’s representations are sensitive to the visual properties implied by the linguistic input.

2.3. Vision-Language Models

We evaluate four different CLIP-based Vision Transformers with different numbers of parameters and training regimes in order to test the generalizability and robustness of implied visual feature effects.

The Vision Transformer (ViT) architecture adapts the Transformer to handle visual data (Dosovitskiy et al., 2021). The ViT divides an image into fixed-size non-overlapping patches that are then linearly embedded into input vectors. A classification head is attached to the output to produce the final prediction. Despite their simplicity and lack of inductive biases (e.g., convolutional layers), ViTs have achieved competitive performance on various visual tasks, especially when pre-trained on large datasets (Dosovitskiy et al., 2021; Schuhmann et al., 2022).

CLIP (Contrastive Language–Image Pre-training) employs contrastive learning to associate images with text descriptions (Radford et al., 2021). The model jointly trains a ViT image encoder and a text encoder to predict the correct pairings of (image, text) pairs. This allows CLIP to learn a shared semantic space between images and text. We evaluate four pre-trained CLIP models:

ViT-B/32: The base model from (Radford et al., 2021). ViT-B/32 uses a patch size of 32px and has 120M parameters. It was trained on 400 million 224x224 pixel image-text pairs over 32 epochs.

ViT-L/14: The best-performing model from (Radford et al., 2021, described in the paper as ViT-L/14@336px). ViT-L/14 uses a patch size of 14px and has 430M parameters. It was pre-trained in the same manner as ViT-B/32 and then fine-tuned at 336px for one additional epoch.

ViT-H/14: A larger model based on the CLIP architecture (Ilharco et al., 2021). ViT-H/14 has 1B parameters and was trained on the LAION 2B dataset for 16 epochs (Schuhmann et al., 2022).

ImageBind: an MLLM that learns a joint embedding across six modalities, including images, text, audio, depth, thermal, and IMU data (Girdhar et al., 2023). Internally, a Transformer architecture is used for all modalities. The image and text encoders are based on the ViT-H/14 model.

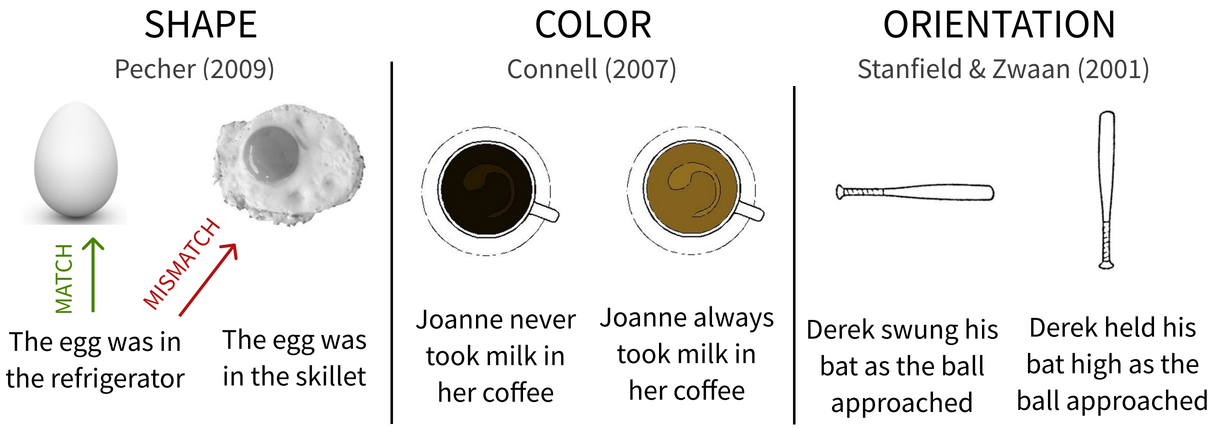
SHAPE
Pecher (2009)
MATCH
MISMATCH
The egg was in the refrigerator
The egg was in the skillet
COLOR
Connell (2007)
Joanne never took milk in her coffee
Joanne always took milk in her coffee
ORIENTATION
Stanfield & Zwaan (2001)
Derek swung his bat as the ball approached
Derek held his bat high as the ball approached

Figure 1: The dataset consisted of pairs of sentences and images, forming quadruplets. Each sentence in a pair implied that an object had a certain visual property (e.g. brown color). Each implied visual property was matched by one of the pair of images. The implied visual properties included SHAPE (**Left,** Pecher et al., 2009), COLOR (**Center**, Connell, 2007), and ORIENTATION (**Right**, Stanfield and Zwaan, 2001).

3. Results

We tested whether MLLMs were sensitive to the implied visual features in the sentence using a t-test. The test compared the probability assigned to images that matched the implied visual features versus those that did not. All of the models, except for the smallest (ViT-B/32), showed a significant effect of SHAPE. ImageBind showed the largest effect: $t(238) = 4.65, p < 0.001$. ViT-B/32 showed an effect in the expected direction but it did not reach significance: $t(238) = 1.81, p = 0.072$.

The results for Color were more varied. Neither the ViT-B/32 and ViT-L/14 models showed a significant effect of match between the color implied by a sentence and the color of an image. Both ViT-H/14 ($t(46) = 2.16, p < 0.05$) and ImageBind ($t(46) = 2.85, p < 0.01$ demonstrated sensitivity to implied color properties although these effects were less robust than for shape.

None of the models showed significant sensitivity to implied ORIENTATION from linguistic cues. The largest numerical effect was shown by ImageBind: $t(94) = 1.09, p = 0.278$ (see Table 3).

Model	Shape	Color	Orientation
ViT-B/32	0.072	0.112	0.965
ViT-L/14	**<0.001**	0.240	0.510
ViT-H/14	**<0.001**	**0.036**	0.323
ImageBind	**<0.001**	**0.006**	0.278

Table 1: p-values from t-tests measuring the effect of matching implied visual features between labels and images. All models except ViT-B/32 show a significant effect for Shape. ViT-H/14 and ImageBind both show significant effects for Color. None of the models show an effect of Orientation.

3.1. Follow-up Analysis of Explicit Features

One potential explanation for the null results reported above is that MLLMs are insensitive to the manipulated visual features like orientation, or that these features are difficult to identify in the image stimuli used. To test this possibility, we ran a follow-up "manipulation check" to determine whether the MLLMs were sensitive to orientation and color when they were explicitly mentioned in the text. The analysis was virtually identical to the primary analysis above, except that we used a sentence template that *explicitly* described specific visual features of the object in question, e.g., "It was a [COLOR] [OBJECT]". We then asked whether the MLLMs could successfully match sentences with explicit visual features (e.g., "It was a red traffic light" vs. "It was a green traffic light").

All models tested showed an effect of both COLOR ($p < .01$) and ORIENTATION ($p < .01$). That is, models assigned higher probability to images with visual features that matched those *explicitly mentioned* in the sentence. This indicates that the MLLMs are sensitive to COLOR and ORIENTATION, and that stimulus quality is sufficient to identify these features.

4. Discussion

Our central question was whether MLLMs showed effects that have been taken as evidence of embodied simulation in humans (Stanfield and Zwaan, 2001). We asked whether MLLMs were sensitive to specific visual features (shape, color, and orientation) that were *implied* but not explicitly mentioned by a verbal description of an event. We found robust evidence of simulation for implied SHAPE, mixed

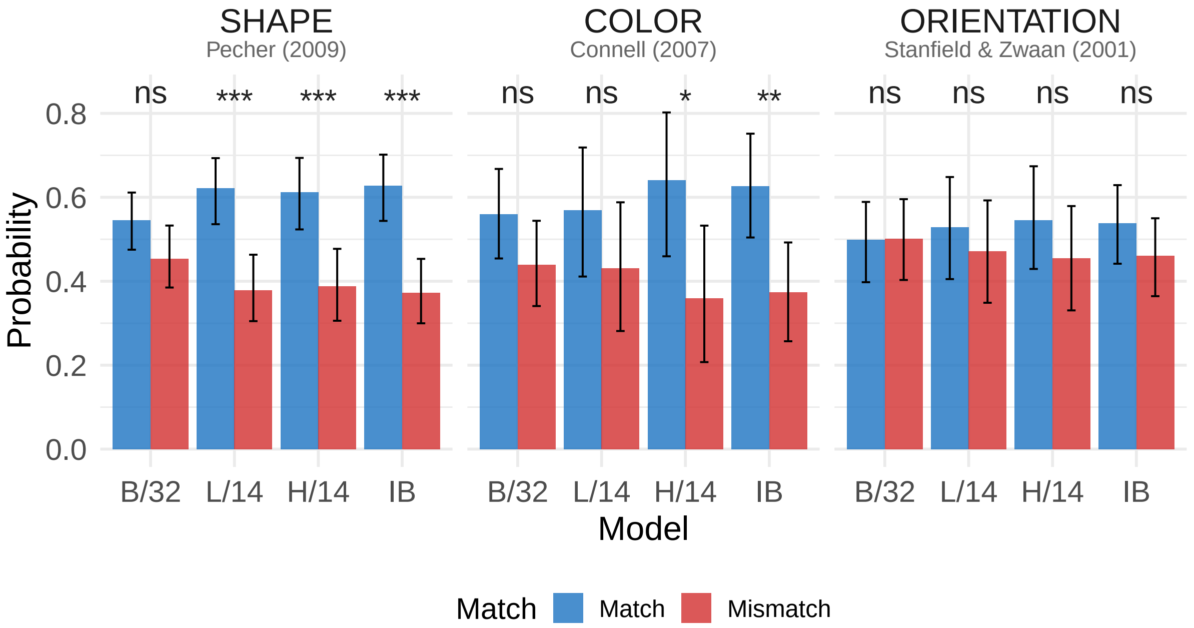

Figure 2: Comparison of mean probability values assigned to images that either matched (blue bars) or did not match (red bars) implied visual features of a sentence. Four Vision Transformer Models (ViT-**B/32**, ViT-**L/14**, ViT-**H/14**, and **I**mage**B**ind)), were evaluated across three datasets (Shape, Orientation, and Color). Error bars denote 95% bootstrapped confidence intervals.

evidence for simulation of implied color, and no evidence of simulation for implied orientation.

Importantly, none of these visual features were explicitly mentioned in the sentences. Thus, if an MLLM exhibits sensitivity to implied shape, it suggests that the model is activating *event-specific* representations of the objects mentioned in a sentence. In humans, an analogous effect is taken as evidence of embodied simulation (Stanfield and Zwaan, 2001; Bergen, 2015). The findings here suggest that such an effect can be produced via exposure to large-scale statistical associations between patterns in images and patterns in text.

It is unclear why MLLMs did not appear to simulate orientation (or color, in some cases). Critically, when either feature was *explicit* in the text, a match effect was obtained (see Section 3.1); this suggests the null effects were not due to overall insensitivity to those visual features. Instead, MLLMs appear to activate some implicit visual features more readily than others. This variation could be driven by noise in the relationship between images and descriptions. Orientation can be influenced by rotation or viewpoints and color similarly varies with lighting. Implicit indications of these features in text labels may therefore be less reliable than indications of more invariant features such as shape. Future work could ask whether color and orientation are *less integrated* with linguistic representations in MLLMs, or simply harder to infer from text descriptions.

Future studies could also explore whether MLLMs simulate modalities beyond vision. There is evidence that humans activate other sensorimotor modalities, such as auditory volume (Winter and Bergen, 2012) and motor action (Fischer and Zwaan, 2008), though evidence for other modalities like olfaction is limited (Speed and Majid, 2018).

Finally, there is considerable debate within psycholinguistics over whether embodied simulation plays a *functional* role in language comprehension, or whether it is epiphenomenal (Ostarek and Bottini, 2021; Mahon and Caramazza, 2008; Glenberg et al., 2008). Future work could contribute to this debate by using MLLMs as “subjects”: specifically, researchers could “lesion” representations of features like shape and ask whether this causally affects processing of sentences implying object shape. This would join the broader “neuroconnectionist” research program that aims to unify research on human cognition and on models inspired by cognition (Doerig et al., 2023).

5. Conclusion

We found that MLLMs are sensitive to whether visual features that are *implied* by a sentence are matched in an image, a phenomenon taken as evidence of embodied simulation in humans.

6. Ethical Considerations and Limitations

The study is limited in that it only evaluates Vision Transformers. Other VLM architectures may produce different associations between text and images. The number of items for some of the datasets was small. Some models may have shown significant match effects with a larger number of items. One potential limitation of the study is that the tasks given to human and LLM participants are not quite analogous. In the picture-verification task, the participant is aware that the implied visual features are irrelevant: their task is to identify whether the object was present in the sentence. The models cannot be so instructed: the measure of association between the sentence and image representations will be based on all features that were useful to the model during CLIP pre-training. Nevertheless, the results show that models are sensitive to these implied features even when they are not explicitly mentioned.

7. Bibliographical References

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Benjamin Bergen. 2015. Embodiment, simulation and meaning. In *The Routledge handbook of semantics*, pages 142–157. Routledge.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Louise Connell. 2007. Representing object colour in language comprehension. *Cognition*, 102(3):476–485.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. 2023. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Martin H Fischer and Rolf A Zwaan. 2008. Embodied language: A review of the role of the motor system in language comprehension. *Quarterly journal of experimental psychology*, 61(6):825–850.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Arthur M Glenberg. 2010. Embodiment as a unifying perspective for psychology. *Wiley interdisciplinary reviews: Cognitive science*, 1(4):586–596.

Arthur M Glenberg, Marc Sato, and Luigi Cattaneo. 2008. Use-induced motor plasticity affects the processing of abstract and concrete language. *Current Biology*, 18(7):R290–R291.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Bradford Z Mahon and Alfonso Caramazza. 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, 102(1-3):59–70.

Dimitri Coelho Mollo and Raphaël Millière. 2023. The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Guillermo Montero-Melis, Jeroen Van Paridon, Markus Ostarek, and Emanuel Bylund. 2022. No evidence for embodiment: The motor system is not needed to keep action verbs in working memory. *cortex*, 150:108–125.

Markus Ostarek and Roberto Bottini. 2021. Towards strong inference in research on embodiment–possibilities and limitations of causal paradigms. *Journal of Cognition*, 4(1).

Diane Pecher, Saskia van Dantzig, Rolf A Zwaan, and René Zeelenberg. 2009. Short article: Language comprehenders retain implied shape and orientation of objects. *Quarterly Journal of Experimental Psychology*, 62(6):1108–1114.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, and Mitchell Wortsman. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Laura J Speed and Asifa Majid. 2018. An exception to mental simulation: No evidence for embodied odor language. *Cognitive Science*, 42(4):1146–1178.

Robert A Stanfield and Rolf A Zwaan. 2001. The effect of implied orientation derived from verbal context on picture recognition. *Psychological science*, 12(2):153–156.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, pages 1–11.

Bodo Winter and Benjamin Bergen. 2012. Language comprehenders represent object distance both visually and auditorily. *Language and Cognition*, 4(1):1–16.

Rolf A. Zwaan and Diane Pecher. 2012. Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts. *PLOS ONE*, 7(12):e51382.