On choosing the vehicles of metaphors without a body: evidence from Large Language Models

Veronica Mangiaterra[1], Chiara Barattieri di San Pietro[1], Federico Frau[1], Valentina Bambini[1], Hamad Al-Azary[2]
[1]Laboratory of Neurolinguistics and Experimental Pragmatics (NEPLab), University School for Advanced Studies IUSS, Pavia, Italy
[2]Lawrence Technological University, Southfield, MI, USA
{veronica.mangiaterra, valentina.bambini}@iusspavia.it, halazary@ltu.edu

Abstract

Since the advent of Large Language Models (LLMs), much work has been devoted to comparing the linguistic abilities of humans and machines. Figurative language, which is known to rely on pragmatic inferential processes as well as lexical-semantic, sensorimotor, and socio-cognitive information, has been often used as a benchmark for this comparison. In the present study, we build on previous behavioral evidence showing that both distributional and sensorimotor variables come into play when people are asked to produce novel and apt metaphors and examine the behavior of LLMs in the same task. We show that, while distributional features still hold a special status, LLMs are insensitive to the sensorimotor aspects of words. This points to the lack of human-like experience-based grounding in LLMs trained on linguistic input only, while offering further support to the multimodality of conceptual knowledge involved in metaphor processes in humans.

1 Introduction

Large language models (LLMs)'s outstanding abilities to solve increasingly complex linguistic tasks (Bubeck et al., 2023; Marvin and Linzen, 2018; Srivastava et al., 2022; Strachan et al., 2024; Webb et al., 2023) have given rise to a theoretical debate on how their skills compare to those of humans (Birhane and McGann, 2024; Mahowald et al., 2024; Mitchell and Krakauer, 2023).

One main point of debate is that, while human linguistic knowledge is acquired through multimodal (sensory, motor, emotional, verbal, etc.) sources (Barsalou, 2008), most LLMs are trained on huge amounts of text only (Bender and Koller, 2020; Bisk et al., 2020), questioning whether LLMs can be considered psychologically valid models of cognition (Bolhuis et al., 2024; Cuskley et al., 2024; Lake and Murphy, 2023).

Pragmatic skills, namely the ability to understand the context-dependent meaning of utterances, including non-literal expressions, have been taken as an ideal test bed to explore the high-level linguistic abilities of LLMs (Barattieri di San Pietro et al., 2023; Liu et al., 2022). Importantly, interpreting non-literal meaning might require skills beyond determining statistical regularities in language. Theoretical pragmatic accounts have highlighted the context-bound nature of pragmatic reasoning (Gumperz, 1982), the need for meta-representation of the speaker's intention (Sperber, 1994), and the potential presence of imagistic processes alongside purely verbal processes (Carston, 2018; Gibbs Jr and Matlock, 2008; Paivio and Walsh, 1993). These claims are supported by neuroimaging experimental evidence that showed, for non-literal meaning processing, activations of brain circuits linked to mental imagery (Mashal et al., 2014) as well as Theory of Mind skills, namely the ability to attribute mental states to others (Bambini et al., 2011; Enrici et al., 2019; Van Ackeren et al., 2012).

Among pragmatic phenomena, figurative language and metaphors are of particular interest. Metaphors are non-literal uses of language that require a context-driven adjustment of the lexically encoded concepts to grasp the interlocutor's intended meaning (Wilson and Carston, 2007). Most studies, including this one, focus on nominal metaphors that involve two terms – a topic and a vehicle – in the form "X is Y". In a metaphor such as "Lawyers are sharks", the concept evoked by the vehicle "sharks" is adjusted by dropping semantic features that are not salient based on the context – e.g., "swims" and "has a fin" – while promoting features leading to the intended meaning, such as "aggressive" and "ruthless". Studies showed that, in addition to lexical-semantic processes, sensorimotor processes also have a role in metaphor processing (Battaglini et al., 2025; Al-

Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 37–44

Azary and Katz, 2021). For example, Al-Azary and Katz (2021) showed that the bodily-action aspects of words are active when processing low-familiar metaphors.

The metaphoric processing abilities of LLMs have been widely investigated (Carenini et al., 2023; Ichien et al., 2024; Neidlein et al., 2020; Prystawski et al., 2022; Wachowiak and Gromann, 2023). LLMs exhibit high accuracy in identifying and interpreting metaphors, yet their performance may rely on different mechanisms compared to humans, as shown by different patterns of errors (Barattieri di San Pietro et al., 2023; Liu et al., 2022) or the need for psychologically informed paradigms to improve their responses (Prystawski et al., 2022).

Compared to metaphor identification and interpretation, metaphor production is a less investigated area, both in humans and machines. Examining how people construct metaphors, and which semantic features guide the process of metaphorical conceptualization, may shed further light on high-level language skills and the role of multimodality, and therefore is a fertile ground for comparison with LLMs.

Katz (1989) investigated metaphor production with a *vehicle selection paradigm*, (i.e., asking participants to produce a metaphor by selecting the metaphor vehicle word among a given set), and found that participants tend to choose concrete words with a moderate semantic distance from the topic. Expanding the study of the semantic features at work in metaphor production, Al-Azary and Katz (2023) investigated the role of semantic richness, namely the amount of semantic information carried by a word (Yap et al., 2012). Specifically, they used two variables reflecting different aspects of semantic richness: Semantic Neighborhood Density (SND) and Body-Object Interaction (BOI). SND was defined as high (dense) or low (sparse), based on the average distance between the word and its semantic neighbors (Buchanan et al., 2001). Body-Object Interaction is a normed variable, derived from human ratings, that indicates the ease with which a human body can physically interact with a word's referent (Siakaluk et al., 2008), thus reflecting sensorimotor richness. Humans are more likely to experience physical interaction with the referent of a high-BOI word such as 'umbrella' rather than that of a low-BOI word such as 'volcano'. They found that participants prefer vehicles with low SND and low BOI, like 'cloud' or 'rainbow' (rather than high-BOI high-SND words like 'pillow'), resulting in metaphors such as 'Boredom is a cloud' or 'Persuasion is a rainbow'. This indicates that humans are sensitive to both sensorimotor and linguistic richness, as expressed by high SND and high BOI, in metaphor production.

In this study, we tested LLMs in the same metaphor production task to examine whether distributional and sensorimotor features of words, which are relevant for human participants, drive metaphor production in LLMs as well, and whether LLMs show the same direction of effects.

2 Methods

We reproduced the behavioral experiments of the study by Al-Azary and Katz (2023) on LLMs by prompting four models within the open-source family of GPT2 models developed by OpenAI (Radford et al., 2019).

2.1 Material

Materials were taken from the original study. The set of stimuli included 36 topics and 48 potential vehicles. Topics were abstract words, as assessed by low concreteness ratings in the Brysbaert et al. (2014) dataset. Moreover, topics were balanced in terms of SND values extracted from the WINDSORS database (Durda and Buchanan, 2008), with half of them with low SND and half with high SND. Vehicles were concrete words, as assessed by high concreteness ratings in the Brysbaert et al. (2014) dataset and were balanced in terms of both SND values (Durda and Buchanan, 2008) and Body-Object Interaction (BOI) values, extracted from the databases by Bennett et al. (2011) and Tillotson et al. (2008), resulting in four combinations of 12 vehicles each. An example of each category of topic and vehicle is provided in Table 1.

2.2 Models

We employed four pre-trained transformer-based language models developed by OpenAI (Radford et al., 2019): GPT2 (124M parameters), GPT2-medium (355M parameter), GPT2-Large (774M parameters), and GPT2-XL (1.5B parameters). These models differ only in architecture scale: the number of layers, hidden dimensions, and parameters increases progressively, while the training data and objectives remain constant. Specifically, these models are trained on *WebText*, a large corpus of English created by scraping 45 million links from

Topic (abstract)	Low SND	Nostalgia
	High SND	Empathy
Vehicle (concrete)	High SND - High BOI	Seed
	High SND - Low BOI	Butterfly
	Low SND - High BOI	Umbrella
	Low SND - Low BOI	Lighthouse

Table 1: Examples of topics and vehicles presented in the experiment and their respective semantic conditions. Note: SND = Semantic Neighborhood Density; BOI = Body-Object Interaction.

Reddit. Differently from larger GPT models, GPT2 models provide access to the probability distributions over strings of words, a way of testing LLMs' capabilities proven to be more reliable than prompting alone (Hu and Levy, 2023).

2.3 Procedure

In the original study, human participants were presented with the topic and the set of 48 vehicles, and they were asked to choose one of the vehicles to create a comprehensible and apt metaphor, resulting in 36 unique metaphors produced. To replicate this experimental paradigm, we prompted the model with the string "TOPIC is a/an" and we collected the likelihood scores of each candidate vehicle. For each topic, we then extracted, the normalized probability distribution over the whole set of vehicles. Probabilities were normalized using a softmax-like transformation with a temperature parameter τ = 0.05 to control sharpness. We randomized vehicle order to reduce ordering bias and used a fixed seed (set prior to execution) to ensure reproducibility. The code was adapted from Carenini et al. (2023). Even if GPT models are not explicitly instructed to produce a metaphor, all possible candidates given the prompt "TOPIC is a/an" form a metaphorical expression. Thus, we expect that the models, when completing the prompt in the most likely way as they are trained to do, will provide us with what the LLM consider the most comprehensible and apt metaphor among the possible metaphorical combinations.

2.4 Statistical Analysis

To test whether LLMs choose vehicles according to the semantic features of both topics and vehicles, we fit a set of Linear Mixed Models using lme4 and lmerTest packages (Bates et al., 2015; Kuznetsova et al., 2017) for each GPT2 model, separately. We consider the z-scaled probability of the vehicle as the dependent variable and Topic SND, Vehicle SND, and Vehicle BOI as interacting categorical predictors. A random intercept was added to account for the variability of individual vehicles. The resulting formula was: *lmer(probability ~ vehicle-SND * vehicle-BOI * topic-SND + (1|vehicle).* Alpha level was set at .05.

3 Results

The fitted linear mixed models showed a main effect of SND of the vehicles for GPT2-Medium (β = 0.64, t = 2.316, p = 0.025), GPT2-Large (β = 0.612, t = 2.264, p = 0.029), and GPT2-XL (β = 0.589, t = 2.159, p = 0.036). These three models showed a higher probability of choosing low-SND vehicles compared to high-SND (Figure 1).

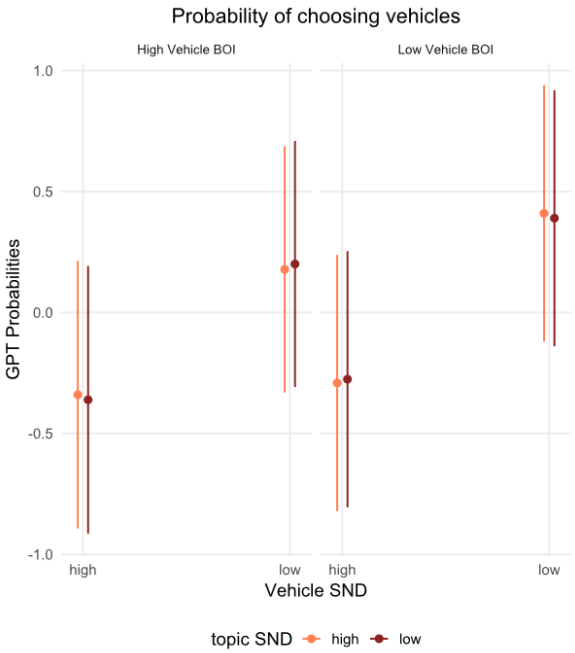

Figure 1: Effect of Vehicle SND on z-scaled probabilities of choosing vehicles extracted from GPT2-Large. Similar patterns are observed in the other GPT2 models.

Concerning GPT2, the linear mixed model showed an interaction between Topic SND and Vehicle SND (β = -0.056, t = -2.321, p = 0.020), meaning that GPT2 assigned higher probabilities to low-SND vehicles when completing metaphors with high-SND topics.

No effect of vehicle BOI was found for all four GPT2 models, regardless of their size. Some examples of the metaphors produced by models (and

by humans from the original Al-Azary and Katz (2023) study) can be found in Appendix A.

4 Discussion

In this work, we examined the role of distributional (SND) and sensorimotor (BOI) variables in metaphor production by LLMs. To do so, we prompted four models of the GPT2 family with a metaphor production task from Al-Azary and Katz (2023). In the original study, they found that human participants, when asked to choose a vehicle to construct an apt metaphor, tend to rely on both SND and BOI, preferring vehicles with low values in both variables. Differently, LLMs are not sensitive to BOI values, equally choosing low- and high-BOI words as vehicles, but showing a preference for low-SND vehicles, as humans do. In other words, LLMs align with human participants in choosing low-SND words such as 'puzzle', yet, they assign similarly high probabilities to metaphors like "Boredom is a vacuum" (high BOI) and "Persuasion is an eagle" (low BOI), showing no sensitivity to BOI values — in contrast to human participants, who preferred metaphors with low-BOI words such as "Boredom is a cloud" and "Persuasion is a butterfly".

In their original study, Al-Azary and Katz (2023) argued in favor of an advantage of low semantic richness in the emergence of metaphorical meaning. In line with that, Al-Azary and Buchanan (2017) showed that semantic richness, operationalized as a greater number of closer semantic neighbors (high SND), hinder the comprehensibility of metaphors, in that the adjustment process to derive the intended meaning in semantically rich concepts is more costly compared to semantically less-rich concepts, as many more features of the word need to be discarded (Kintsch, 2000), hence leading to a preference for low semantic richness (confirmed in studies on literary metaphors (Reid et al., 2023; Mangiaterra et al., 2024)).

Similarly, LLMs seem to adhere to the tendency toward low semantic richness exhibited by humans both in comprehension (Al-Azary and Buchanan, 2017) and production (Al-Azary and Katz, 2023). However, LLMs do so by relying only on the distributional features of words (SND) rather than their sensorimotor content (BOI). This pattern did not change across the four GPT2 models employed in this study, suggesting that the ability to rely on sensorimotor-relevant features is not significantly enhanced by simply model scale. Although we did not find evidence of sensorimotor effects LLMs in this task, we cannot rule out sensorimotor effects in other tasks, which can be an area of future research. However, the lack of ability of the four GPT2 models to modulate experience-based aspects of vehicles is in line with previous work tackling (embodied) cognition in language models. Xu et al. (2023) found that, while language models have human-like representations of words in non-sensorimotor domains, they do not align with respect to words related to sensory and motor domains. Similarly, Lee et al. (2025) showed that even if models could approximate human perceptual ratings, they do so by relying more on linguistic cues. Even multimodal LLMs fall behind human performance in classical tasks exploring the activation of implied sensorimotor features of concepts in language processing (Jones et al., 2024). There is also specific evidence of a poor handling of sensorimotor experience in LLM's processing of metaphor. For instance, Barattieri di San Pietro et al. (2023) found that ChatGPT performs better on metaphors expressing a psychological characteristic of the topic compared to metaphors capitalizing on physical features. All of these studies employed larger models compared to the GPT2 models used in the present work, confirming that even if scaling improves language performance, some sensorimotor aspects cannot fully emerge from training on textual input or input from visual modality only.

The role of both disembodied (SND) and embodied (BOI) sources of linguistic knowledge in metaphor production in humans (Al-Azary and Katz, 2023) supports the theoretical account of the "dual coding theory" (Paivio, 1979), namely the perspective according to which the nature of concepts is both symbolic and imagistic. Our results on LLMs may provide evidence in support of these claims about the nature of conceptual representation, as well. On the one hand, GPT2 models have access to verbal input, and this is reflected in their human-like behavior with respect to distributional aspects of words, meaning that a certain portion of linguistic abilities may have a purely disembodied nature and can be acquired with exposure to text only. On the other hand, a certain part of linguistic skills in humans is linked to sensorimotor experience, and the necessity of this experience for human-like linguistic behavior is reflected in the absence of these aspects in language models, which lack this source of knowledge. This implies that

even huge amounts of linguistic input cannot replace the multimodal sources from which linguistic knowledge is acquired in humans.

A suggestive hypothesis arising from the data is that LLMs may behave in a way similar to humans with low imagery and high vocabulary skills. Interestingly, electrophysiological and behavioral studies accounting for individual differences reported that different profiles emerge in metaphor processing (Battaglini et al., 2025) and that a greater reliance on the semantic/distributional route may also be present in humans. Shen et al. (2015) found that participants with low mental imagery abilities showed a greater neurophysiological response linked to semantic mismatch and no imagery activations when processing metaphors, while high-vocabulary participants are found to be less sensitive to sensorimotor features of words (Frau et al., 2025) but more sensitive to their semantic neighbors (Pexman and Yap, 2018).

This work confirms the importance of the study of figurative language and metaphorical abstraction to disentangle the subtle aspects of linguistic processing that distinguish between a formal and a functional human-like competence of language in large language models (Mahowald et al., 2024). The evidence reported in this study may add to the broader debate on which aspects of human language abilities LLMs are actually modeling (Dove, 2024). In particular, our results suggest that LLMs may be a valid model in cognitive science for those linguistic aspects relying on distributional features, while their limitations should be taken into account when considering language skills that require a multimodal conceptual representation.

5 Limitations

First of all, in the experiment, we chose to employ GPT2 models and, even if scaling does not necessarily improve human-like grounding (see Discussion), larger or multimodal models could provide different results for our metaphor production task. Moreover, all of the vehicles used in the task are concrete words and thus they potentially carry sensorimotor information, regardless of their BOI differences. So, the lack of sensitivity to this fine-grained feature does not exclude the possibility of a more general concreteness effect as found in Katz (1989). In addition, the use of only concrete words as vehicles did not allow us to explore the role of other experiential variables typically found in abstract words (e.g., emotional and interocepective properties) that may contribute to the processes at work in metaphor production for humans and machines.

6 Acknowledgments

This work received support from the European Research Council under the EU's Horizon Europe program, ERC Consolidator Grant "PROcessing MEtaphors: Neurochronometry, Acquisition and Decay, PROMENADE" [101045733]. The content of this article is the sole responsibility of the authors. The European Commission or its services cannot be held responsible for any use that may be made of the information it contains.

References

Hamad Al-Azary and Lori Buchanan. 2017. Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory & Cognition*, 45:296–307.

Hamad Al-Azary and Albert N Katz. 2021. Do metaphorical sharks bite? simulation and abstraction in metaphor processing. *Memory & Cognition*, 49:557–570.

Hamad Al-Azary and Albert N Katz. 2023. On choosing the vehicles of metaphors 2.0: the interactive effects of semantic neighborhood density and body-object interaction on metaphor production. *Frontiers in Psychology*, 14:1216561.

Valentina Bambini, Claudio Gentili, Emiliano Ricciardi, Pier Marco Bertinetto, and Pietro Pietrini. 2011. Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain research bulletin*, 86(3-4):203–216.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Chiara Battaglini, Federico Frau, Veronica Mangiaterra, Luca Bischetti, Paolo Canal, and Valentina Bambini. 2025. Imagers and mentalizers: capturing individual variation in metaphor interpretation via intersubject representational dissimilarity. *Accepted at Lingue e Linguaggio*. Preprint: https://osf.io/n8x36.

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.

Stephen DR Bennett, A Nicole Burnett, Paul D Siakaluk, and Penny M Pexman. 2011. Imageability and body–object interaction ratings for 599 multisyllabic nouns. *Behavior research methods*, 43:1100–1109.

Abeba Birhane and Marek McGann. 2024. Large models of what? mistaking engineering achievements for human linguistic agency. *Language Sciences*, 106:101672.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, and 1 others. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. Three reasons why ai doesn't model human language. *Nature*, 627(8004):489.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Lori Buchanan, Chris Westbury, and Curt Burgess. 2001. Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, 8(3):531–544.

Gaia Carenini, Louis Bodot, Luca Bischetti, Walter Schaeken, and Valentina Bambini. 2023. Large language models behave (almost) as rational speech actors: Insights from metaphor understanding. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.

Robyn Carston. 2018. Figurative language, mental imagery, and pragmatics. *Metaphor and Symbol*, 33(3):198–217.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Guy Dove. 2024. Symbol ungrounding: what the successes (and failures) of large language models reveal about human cognition. *Philosophical Transactions B*, 379(1911):20230149.

Kevin Durda and Lori Buchanan. 2008. Windsor: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40(3):705–712.

Ivan Enrici, Bruno G Bara, and Mauro Adenzato. 2019. Theory of mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature of human communication. *Pragmatics & Cognition*, 26(1):5–38.

Federico Frau, Luca Bischetti, Lorenzo Campidelli, Elisabetta Tonini, Emiko J Muraki, Penny M Pexman, and Valentina Bambini. 2025. Understanding with the body? testing the role of verb relative embodiment across tasks at the interface of language and memory. *Journal of Memory and Language*, 140:104566.

Raymond W Gibbs Jr and Teenie Matlock. 2008. Metaphor, imagination, and simulation: Psycholinguistic evidence. In Jr. R. W. Gibbs, editor, *The Cambridge handbook of metaphor and thought*, page 161–176. Cambridge University Press.

John J Gumperz. 1982. *Discourse strategies*. Cambridge University Press.

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations. *arXiv preprint arXiv:2305.13264*.

Nicholas Ichien, Dušan Stamenković, and Keith J Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

Cameron R Jones, Benjamin Bergen, and Sean Trott. 2024. Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, 50(4):1415–1440.

Albert N Katz. 1989. On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and language*, 28(4):486–499.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic bulletin & review*, 7(2):257–266.

Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.

Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.

Jonghyun Lee, Dojun Park, Jiwoo Lee, Hoekeon Choi, and Sung-Eun Lee. 2025. Exploring multimodal perception in large language models through perceptual strength ratings. *arXiv preprint arXiv:2503.06980*.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Veronica Mangiaterra, Chiara Barattieri di San Pietro, and Valentina Bambini. 2024. Temporal word embeddings in the study of metaphor change over time and across genres: a proof-of-concept study on english. In *10th Italian Conference on Computational Linguistics, CLiC-it 2024*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Nira Mashal, Tali Vishne, and Nathaniel Laor. 2014. The role of the precuneus in metaphor comprehension: evidence from an fmri study in people with schizophrenia and healthy participants. *Frontiers in human neuroscience*, 8:818.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736.

Allan Paivio. 1979. *Imagery and verbal processes*. Psychology Press.

Allan Paivio and Mary Walsh. 1993. Psychological processes in metaphor comprehension and memory. In Andrew Ortony, editor, *Metaphor and Thought*, pages 2–307. Cambridge University Press.

Penny M Pexman and Melvin J Yap. 2018. Individual differences in semantic processing: Insights from the calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(7):1091.

Ben Prystawski, Paul Thibodeau, Christopher Potts, and Noah D Goodman. 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nick Reid, Hamad Al-Azary, and Albert N Katz. 2023. Cognitive factors related to metaphor goodness in poetic and non-literary metaphor. *Metaphor and Symbol*, 38(2):130–148.

Zih-Yu Shen, Yi-Ting Tsai, and Chia-Lin Lee. 2015. Joint influence of metaphor familiarity and mental imagery ability on action metaphor comprehension: An event-related potential study. *Language and Linguistics*, 16(4):615–637.

Paul D Siakaluk, Penny M Pexman, Christopher R Sears, Kim Wilson, Keri Locheed, and William J Owen. 2008. The benefits of sensorimotor knowledge: Body–object interaction facilitates semantic processing. *Cognitive Science*, 32(3):591–605.

Dan Sperber. 1994. Understanding verbal understanding. *What is intelligence*, 179:98.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Sherri M Tillotson, Paul D Siakaluk, and Penny M Pexman. 2008. Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, 40(4):1075–1078.

Markus J Van Ackeren, Daniel Casasanto, Harold Bekkering, Peter Hagoort, and Shirley-Ann Rueschemeyer. 2012. Pragmatics in action: indirect requests engage theory of mind areas and the cortical motor network. *Journal of cognitive neuroscience*, 24(11):2237–2247.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Deirdre Wilson and Robyn Carston. 2007. A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts, editor, *Pragmatics*, page 230–259. Palgrave Macmillan, Basingstoke.

Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2023. Does conceptual representation require embodiment? insights from large language models. *arXiv preprint arXiv:2305.19103*.

Melvin J Yap, Penny M Pexman, Michele Wellsby, Ian S Hargreaves, and Mark J Huff. 2012. An abundance of riches: Cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in human neuroscience*, 6:72.

A Appendix

Examples of metaphors with higher probability assigned by GPT2
Boredom is a vacuum (High BOI -Low SND)
Persuasion is an eagle (Low BOI – Low SND)
Prestige is a tiger (Low BOI – Low SND)
Destiny is a bicycle (High BOI – High SND)
Narcissism is a lighthouse (Low BOI – Low SND)
Sadness is a puzzle (High BOI – Low SND)

Examples of metaphors produced by humans
Boredom is a desert (Low BOI – Low SND)
Persuasion is a butterfly (Low BOI – Low SND)
Prestige is a rainbow (Low BOI – Low SND)
Destiny is a seed (High BOI – Low SND)
Narcissism is a volcano (Low Boi- High SND)
Sadness is a cloud (Low BOI – Low SND)