

Exploring Spatial Schema Intuitions in Large Language and Vision Models

Philipp Wicke*

Ludwig-Maximilian-University, Munich
Munich Center for Machine Learning (MCML)
pwicke@cis.lmu.de

Lennart Wachowiak*

King's College London
Imperial College London
lennart.wachowiak@kcl.ac.uk

Abstract

Despite the ubiquity of large language models (LLMs) in AI research, the question of embodiment in LLMs remains underexplored, distinguishing them from embodied systems in robotics where sensory perception directly informs physical action. Our investigation navigates the intriguing terrain of whether LLMs, despite their non-embodied nature, effectively capture implicit human intuitions about fundamental, spatial building blocks of language. We employ insights from spatial cognitive foundations developed through early sensorimotor experiences, guiding our exploration through the reproduction of three psycholinguistic experiments. Surprisingly, correlations between model outputs and human responses emerge, revealing adaptability without a tangible connection to embodied experiences. Notable distinctions include polarized language model responses and reduced correlations in vision language models. This research contributes to a nuanced understanding of the interplay between language, spatial experiences, and the computations made by large language models.¹

1 Introduction

Large language models (LLMs) excel in varied NLP tasks like text generation, sentiment analysis, or summarization. Nonetheless, an underexplored facet in the study of LLMs pertains to the concept of embodiment. Unlike embodied systems in robotics, where the physical form plays a central role in shaping the system's abilities, LLMs lack a direct connection between sensory perception and physical action. Within this context, we investigate the extent to which LLMs, despite their lack of direct embodiment, might capture the implicit, often sensory-derived, conceptual structures that underlie human language and cognition. In our analysis, we

* Both authors contributed equally.

¹Project site: https://cisnlp.github.io/Spatial_Schemas/

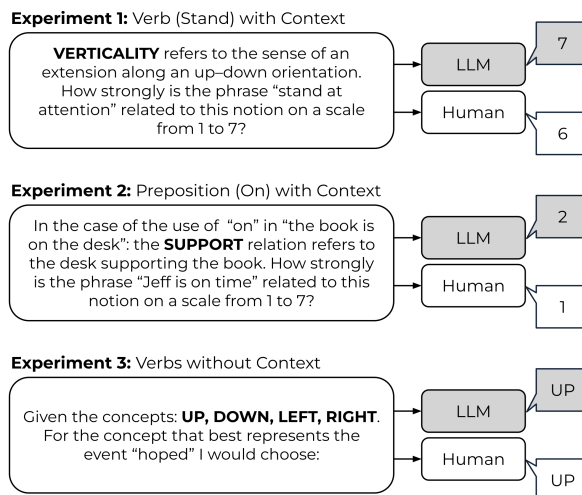


Figure 1: Overview of the three experiments

make use of image schema theory, which provides a set of spatio-temporal cognitive building blocks that are learned in early infancy based on reoccurring sensorimotor experiences (Lakoff and Johnson, 2008; Barsalou, 2008; Johnson, 2013). For example, the image schema SUPPORT is learned by observing and experiencing objects like tables or chairs supporting other objects. According to theory, the same image schema is reused to structure our language and thought, even in regard to abstract topics. For instance, when talking about emotional support, we say *to support a friend*.

Given that LLMs lack grounding, our research centers around a fundamental question: Can LLMs encode people's intuitions about the image schematic basis of words and phrases? In the subsequent sections, we present the rationale behind our inquiry and the methodologies for investigating the intricate relationship between language, embodiment, and the core aspects of human cognition. In summary, we make the following contributions:

- We use LLMs and vision language models (VLMs) to reproduce **three psycholinguistic experiments** (Fig. 1) that connect language to humans' spatial intuitions

- We find that, in many instances, the answers of the largest models and of human participants **show moderate to strong correlations** even though the model’s language use is not grounded in an embodied or enacted sense
- Crucial differences to human answers remain: the output of small and base models often shows low correlations, LLMs tend to give **polarized responses** (selecting either 1 or 7 on a scale), and VLMs show lowest correlations with open-source VLMs showing none.

2 Background

2.1 Image Schemas in Cognition & Language

Image schemas, rooted in cognitive linguistics and embodied cognition, have emerged as a foundational concept in our understanding of human language and cognition. These dynamic mental structures, originally proposed by [Lakoff and Johnson \(2008\)](#), serve as the cognitive building blocks for our conceptualization of the world. Image schemas are pre-linguistic and perceptual in nature, providing a fundamental means for humans to ground abstract concepts in concrete sensory and motor experiences ([Johnson, 2013](#); [Barsalou, 2008](#)). This grounding is pivotal in comprehending and communicating complex ideas, as it bridges the gap between sensorimotor experiences and the vast array of abstract concepts that are integral to human thought and language. The centrality of image schemas in cognitive processes underscores their profound influence on language, from shaping our metaphors and linguistic expressions to facilitating our ability to reason, plan, and understand the world ([Di Paolo et al., 2018](#); [Hampe, 2005](#)). An example of an image schema is VERTICALITY. We can physically experience the image schema VERTICALITY by standing upright or seeing one object positioned above another. In turn, these learned schemas can help us comprehend and communicate abstract concepts such as complex emotions (*I feel down*) or power dynamics (*she ranks high*).

Evidence for image schemas is provided by various psycholinguistic studies ([Mandler, 1992](#); [Gibbs et al., 1994](#); [Boroditsky, 2000](#); [Richardson et al., 2001](#); [Gibbs, 2005](#)). For example, [Richardson et al. \(2001\)](#) has participants pick one of four arrows ($\uparrow, \downarrow, \leftarrow, \rightarrow$) that best represent a concrete or abstract action on horizontal or vertical dimension (e.g. concrete, horizontal: pushed / abstract, ver-

tical: obeyed). The results indicate a common correlation that points towards underlying schemas in language. But how are these schematic intuitions represented by computational models of language?

2.2 Image Schemas in Language Models

LLMs are trained on vast amounts of text and code in order to construct a model of language that can be used for a variety of NLP tasks without having been specifically trained on these tasks ([Radford et al., 2019](#)). The ability of these emergent properties scales with the size of these models ([Wei et al., 2022](#)). Moreover, VLMs, e.g. GPT-4 ([OpenAI, 2023](#)), are trained on images, text and code. Yet, neither LLMs nor VLMs are embodied systems in the sense that they never connect “perception to action directly” ([Brooks, 1991](#)). This leads to the symbol grounding problem, as posed by [Harnad \(1990\)](#), questioning whether symbols can derive meaning just from other symbols alone (as would be the case for text-based LLMs) or whether they would need to be connected in a bottom-up fashion to sensory representations. To this day, such questions are being critically discussed in the NLP community ([Bender and Koller, 2020](#)). Research with embodied computational systems (e.g. robots) often works with the implicit premise (or challenge) that a system’s physical form contributes to its technical capabilities, its affordances ([Brohan et al., 2023](#)). One way to look at the effect of embodiment in LLMs is provided by [Wicke \(2023\)](#), who shows that the degree of perceived embodiment of an action word can have a positive effect on an LLM’s capability to interpret figurative language.

3 Related Work

3.1 Human Behavioral Experiments with Large Language Models

Using LLMs as human stand-in participants for psychology experiments has recently gained attention ([Futrell et al., 2019](#); [Linzen and Baroni, 2021](#); [Dillion et al., 2023](#); [Harding et al., 2023](#); [Aher et al., 2023](#)). Such use can be motivated by wanting to generate initial hypotheses for an experiment, pilot a new design, and gain insight into human cognition based on the assumption that LLMs trained on a large amount of human-generated text will produce similar output to that of human participants ([Dillion et al., 2023](#)). For instance, [Dillion et al. \(2023\)](#), who propose such a use, report a high correlation of 0.95 between human answers and GPT-3.5

answers on moral judgment tasks. At the same time, they acknowledge that current LLMs are bad at capturing variation and diversity present in human responses and are biased towards responses of people from certain countries, economic backgrounds, genders, etc. [Harding et al. \(2023\)](#) critique the use of LLMs to replace human participants and question the informativeness of the LLM’s output.

Another motivation to simulate psychological experiments with LLMs is to gain insights not into human cognition but into the capabilities of language models themselves. Reproducing various experiments with LLMs, one can compare the LLM output with how humans behaved in the real experiment, thereby establishing the “human-likeness” of the model’s text generations. The usefulness of such experiments has also been suggested with respect to psycholinguistics, where experiments can show what properties of language can be successfully processed, reproduced, or generated by LLMs ([Houghton et al., 2023](#)).

A survey by ([Linzen and Baroni, 2021](#)) presents studies of neural networks’ syntactic abilities and their broader implications for linguistic theory. [Dentella et al. \(2023\)](#), for example, show that LLMs fail at distinguishing grammatical and ungrammatical sentences in a similar way to people. [Futrell et al. \(2019\)](#) test four neural network language models on artificial sentences with syntactically complex structures (subordinate clauses and the Garden Path effect) to analyze their syntactic representations. Their findings indicate that LSTMs trained on large datasets represent syntactic states comparably to an RNN trained on a small dataset, while an LSTM trained on a small dataset performs poorly or only weakly. Other studies find that on many psychology tasks, the LLM output is comparable to human answers, even showing similar cognitive biases ([Hagendorff et al., 2022](#); [Dasgupta et al., 2022](#)). [Hagendorff et al. \(2022\)](#) show that these cognitive biases tend to vanish when experimenting with the most recent models, such as ChatGPT and GPT-4. [Aher et al. \(2023\)](#) extend the idea of repeating prominent experiments with LLMs. Specifically, they not only look at a single output of an LLM given some experiment prompt but try to simulate different demographics by prompting the model multiple times with different personas attached to each prompt. Generally, one needs to be aware that the same or similar experiments might have been in the training data.

3.2 NLP for Image Schemas

LLMs have only sparingly been used for image schema-related tasks. Initial research on the computational processing of image schemas was restricted to spectral cluster analysis ([Gromann and Hedblom, 2017](#)), whereas more recent work ([Wachowiak et al., 2022](#); [Wachowiak and Gromann, 2022](#)) uses the language model XLM-RoBERTa ([Conneau et al., 2020](#)). [Wachowiak and Gromann \(2022\)](#) show that language models can be fine-tuned to classify sentences based on eight image schema classes, with an accuracy between 57% and 80% depending on the language of the data. Although their model requires first seeing more than 1,000 correctly annotated samples, these results indicate that, in principle, it is possible for a neural model trained only on text to pick up the pattern indicating a specific image schema in natural language. However, in contrast to the experiments in our work, the language samples they use are manually collated or created by experts. The psycho-linguistic experiments that build the foundation for the present work take a different approach, letting multiple people annotate phrases based on intuitions and felt relatedness. Accordingly, the expert-annotated image schemas are annotated with a single discrete label per sample, while the psycho-linguistic experiments inspiring this study lead to fine-grained annotations using ordinal scales for five different image schemas per phrase. [Kamath et al. \(2023\)](#) test various VLMs for whether they can correctly classify simple spatial configurations, such as A being left, right, under, or over B. Thus, they test for image schematic relations in their original physical form rather than the abstract extensions. They find models to perform poorly on the task, with the best model only achieving an accuracy of 60%, while humans achieve 99%, one of the reasons being that prepositions occur infrequently and in an ambiguous manner in the training data. Similar work by [Jassim et al. \(2023\)](#) shows shortcomings of VLMs’ understanding of spatial configurations and intuitive physics when given different visual inputs of simulated spatial scenes.

From the corpus of related works delving into psycholinguistic studies probing spatial schema intuitions, we selected three experiments for comparison, each exhibiting variations in word class and context. Experiment 1 ([Gibbs et al., 1994](#))

Model Type	Model Name	Model Size	API Endpoint	Open Source
LLM	GPT-3 _{base}	175b	davinci-003	-
LLM	GPT-3 _{inst}	N/A	text-davinci-003	-
LLM	GPT-4	N/A	gpt-4-0613	-
LLM	LLaMA-2-13b	13b	Llama-2-13b-chat-hf	✓
LLM	LLaMA-2-70b	70b	Llama-2-70b-chat-hf	✓
VLM	GPT-4 _{vision}	N/A	gpt-4-vision-preview	-
VLM	IDEFICS-80b _{base}	80b	idefics-80b	✓
VLM	IDEFICS-80b _{inst}	80b	idefics-80b-instruct	✓

Table 1: Models selected for experiments. LLM: large language model, VLM: vision–language model.

showcases a verb (stand) within context, Experiment 2 (Beitel et al., 2001) features a preposition (on) within context, and Experiment 3 (Richardson et al., 2001) employs different verbs without context (see Fig. 1). Our rationale for excluding other studies stemmed from their failure to offer sufficient variation or to provide access to their original questionnaire. Moreover, the chosen experiments span a spectrum of distinct spatial schemas.

4 Method

This section provides an overview of the general model and prompt selection, while following subsections detail the setup for individual experiments.

Model Selection. We recreate the experiments using closed and open-source language models. The open-source LLaMA-2 instruction-tuned models (Touvron et al., 2023) are chosen in two sizes²: LLaMA-2-13b and LLaMA-2-70b. The closed source models include: GPT-3 base (davinci-002), GPT-3.5 instruction-tuned (text-davinci-003) and GPT-4 (gpt-4-0613). Instruction-tuned versions have been chosen over base models because instruction-tuned models tend to perform better (Touvron et al., 2023). This difference can be observed across all of our experiments in varying degrees. For the third experiment, we include three VLMs: the two open-source models IDEFICS-80b (Laurençon et al., 2023) in base and instruction-tuned variants, as well as the GPT-4 Vision model (gpt-4-vision-preview).

Prompt Selection. Regarding their format, we keep the prompts as close to each original experiment as possible. Given the information provided in each paper, we reuse the wording of the image

²Results of further experiments with non-instruct LLaMA models and their 7b version are included in the repository.

schema definitions and the items being evaluated. Besides these, we write our own instructive sentence that prompts the LLM to rate each item since the original instructions given to the human participants are unfortunately not provided in any of the papers. As suggested by Aher et al. (2023), we optimize the instructive sentences by choosing a sentence that maximizes the fraction of valid model answers for each task. In experiments 1 and 2, valid answers consist of the numbers 1 to 7, while in experiment 3, the valid answers consist of the four possible directions. Given a set of valid answers V and a prompt k , the validity score is computed as:

$$\sum_{a \in V} p(a|k) \quad (1)$$

This way of finding a prompt allows us to get valid answers by only looking at their form but not at their content. Thus, we adopt this method that prevents overfitting caused by prompt-engineering.

Recreating a prompt that closely mirrors the approach of the original paper would involve consolidating all the stimuli into a single list and instructing the models to rate each stimulus in combination with each image schema. We tried conditions in which the model had to rate all stimuli for all image schemas and all stimuli for a single schema. Additionally, we tried averaging multiple of those runs based on different stimuli orders. However, all preliminary experiments revealed that this comprehensive prompt yielded impractical or even unmeaningful responses. Frequently, models redundantly reproduced identical outputs for each item in the list. Consequently, we opted to refine our approach by providing the models with a single stimulus per input prompt.

Evaluation. To evaluate how well the models can predict the human participant’s judgment, we compute the Spearman correlation coefficient (Spearman, 1904) between the human and the LLM ratings for each image schema per experiment. For interpretation, we use the labels weak, moderate (>0.4) and strong (>0.7), common in psychology literature (Akoglu, 2018). In each of the three independent experiments, we addressed the challenge of multiple testing by employing the Benjamini-Hochberg correction for False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). This correction was applied to account for the evaluation of multiple correlations within each experiment, ensuring a controlled rate of false positives and bol-

stering the statistical reliability of our findings. We use the *statsmodels* python package (Seabold and Perktold, 2010) to evaluate the corrected p-values. All reported p-values are corrected and marked with * for $p < 0.05$ as statistically significant.

4.1 Experiment 1 - Gibbs et al. (1994)

The Original Experiment. Firstly, we reproduce experiments by Gibbs et al. (1994), testing people’s intuitions about the image schemas that underlie various uses of the verb *to stand*. Given 32 phrases and the definitions of five relevant image schemas, they asked 27 participants to rate the relatedness between each image schema and each phrase on a Likert scale from 1 (“not at all related”) to 7 (“very strongly related”). The image schemas used in this experiment are: BALANCE, VERTICALITY, CENTER–PERIPHERY, RESISTANCE, and LINKAGE. Participants have to rate the relatedness of all 32 phrases to a single image schema before moving to the next (we refer to this data as $Gibbs_{stand}$). Before giving the rating for a particular image schema, the schema is introduced with a short definition. For instance, VERTICALITY is introduced as referring “to the sense of an extension along an up–down orientation”. The order in which the image schemas have to be rated is counterbalanced using five different orders overall. Additionally, Gibbs et al. repeat the experiment using the same 32 phrases but with *stand* being replaced by a word with a synonymous meaning (we refer to this data as $Gibbs_{syn}$). For example, “to stand the test of time” is substituted by “to pass the test of time”.

The LLM Experiment. To extract ratings from an LLM, we retrieve the most likely answer generated, i.e. the answer received with a *temperature* of 0 in the OpenAI API or a *top_k* of 1 with HuggingFace. Alternatively, one could consider the probability for each valid answer, the numbers between 1 and 7, and compute the sum of each number weighted by their likelihood, normalized by the sum of all seven numbers’ likelihoods. While this takes the LLM’s uncertainty into account, it also requires seven times the amount of compute compared to simply taking the top answer. As results only varied minimally, we chose the cost-effective methodology of selecting only the top answer.

Prompting. For experimenting, we started with a basic input text as close as possible to the original experiment, for example:

Consider the notion of VERTICALITY. Verticality refers to the sense of an extension along an up–down orientation. How strongly is the phrase “stand at attention” related to this notion on a scale from 1 (not at all related) to 7 (very strongly related)?

The different image schema definitions used in these prompts can be found in Appendix A1. Given such a text as the start of the input, we try to find a way to end the prompt so that the model’s output probabilities for the next token converge towards 100% when summed for all valid 7 numbers. This final bit of the prompt depends on the model. Our prompt ending choices are described in Section 5.1, and all used phrases are listed in the Appendix A3.

4.2 Experiment 2 - Beitel et al. (2001)

The Original Experiment. Beitel et al. (2001) repeat the experimental paradigm established by Gibbs et al. (1994), however, with a new set of phrases, all containing the preposition *on*. Given the focus on the word *on* instead of *to stand*, they also select a different set of image schemas that are more relevant in this case, namely: SUPPORT, PRESSURE, CONSTRAINT, COVERING, VISIBILITY. Instead of having access to a general definition of each image schema, the participants can now check an example sentence for which five introduction statements explain how each image schema relates to it. For example, the experiment introduces SUPPORT by saying, “In the case of the use of “on” in the “the book is on the desk”: the SUPPORT relation refers to the desk supporting the book”. A list of introductions for all five image schemas is available in the Appendix Table A2.

The LLM Experiment. Given its identical structure, the experiment is conducted in the same manner as with the data from Gibbs et al. (1994).

4.3 Experiment 3 - Richardson et al. (2001)

The Original Experiment. Richardson et al. (2001) provide experimental evidence for image schemas via two different experiments. The first experiment presents human participants with lists of 30 verbs ranked by concreteness (based on the MRC psycholinguistic database (Coltheart, 1981)). The verbs are represented using an agent-patient relation through a circle-square depiction, e.g. \bigcirc *offended* \square . Distinguished by abstractness versus concreteness of the verbs, those verbs were further divided into three groups based on their primary direction: horizontal, vertical and neutral. This

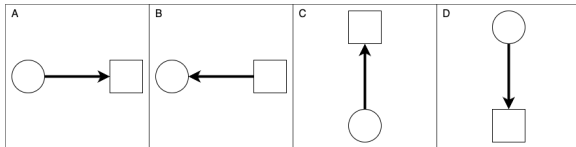


Figure 2: Target images from the original study by Richardson et al. (2001). Each participant was asked to match 30 verbs to one of the images (A-D).

resulted in a 2x3 factor design of concreteness by directionality. The task for each participant was to choose one of four images that best represents the action. The images use the same circle-square relation, but with an arrow representing the directionality on the horizontal and vertical axis (\leftarrow , \rightarrow , \uparrow , \downarrow). The choice of images is depicted in Figure 2. As an example, for the horizontal, concrete item \bigcirc *offended* \square the participant needs to select one of the A-D images (Fig. 2). It is important to note that Richardson et al. (2001) analyse the results with respect to the primary direction (horizontal/vertical). The second experiment by Richardson et al. (2001) requires participants to draw a schematic representation of items from the first experiment. Since we do not include visual generation, we do reproduce the second experiment in the presented work.

The LLM Experiment. The original experiment involved the use of four visual depictions (see Fig.2) alongside a list of verbs. In this setup, participants were exposed to all the images and words simultaneously. The word presentation order was randomised, and the images were labeled A-D. To replicate the experiment, which necessitates a visual input, we also employ VLMs. In order to enable comparisons with our selected (text-only) LLMs, we design a tripartite experimentation.

In the first phase, we opted to translate the visual depictions into **textual** representations of the underlying schemas, specifically the words *up*, *down*, *left*, and *right*. For the second phase, we conducted the experiment again, but this time with **pseudo-visual** renderings using Unicode arrows (\uparrow , \downarrow , \leftarrow , \rightarrow). In the final phase, we turned to a limited selection of VLMs to rerun the experiment with textual and **visual** input, showing the actual images from Fig. 2. This tripartite approach allows us to investigate the impact of visual input on language models, compare it to text-only models, and explore the utility of different forms of visual and textual representations across three conditions.

Importantly, we compare the model responses

with the results from the Richardson et al. (2001) experiment. Initially, our experiments included the analysis of left, right, up and down decisions, which require an additional encoding of the agent-patient relation through a circle-square depiction. This encoding diffused any correlations and we identified that a restrictions to primary directions (horizontal/vertical) by grouping up/down and left/right options in the analysis yields a better reflection of the the models correlations, since it does not require an agent-patient encoding.

Prompting. Detailed lists of prompts for each condition can be found in Appendix Table A5. We observed a significant influence on the model’s response based on the order in which the four options were presented, whether as concept words, Unicode arrows or images. This is in-line with findings by Pezeshkpour and Hruschka (2023). Regardless of different option orders (e.g., \uparrow , \downarrow , \leftarrow , \rightarrow), the model most often favoured the first option (e.g., \uparrow) as a default answer for most items. To mitigate this effect and to acquire a distribution for each verb, we conducted all 24 possible permutations of the choices per word (4! permutations of 30 verbs).

VLMs require a label attached to each image, which allows them to formulate a choice, e.g. “Image A”. We observed that labels with an implicit (alphabetical or numerical) order, e.g. A, B, C or 1, 2, 3, introduce a selection bias. Hence, we use arbitrary labels (VMBR, WJZX, XQHL, YGPK).

Lastly, the GPT-4 models are a commercial product that include safety guardrails in their system prompt in order to avoid hate speech, abuse or disinformation. Often, GPT-4 models would tend to refuse to answer subjective questions. Consequently, we modified instructions for GPT-4 in order to force subjective model answers (Fig. A5).

5 Results

5.1 Experiment 1 — Gibbs et al. (1994)

Prompt Selection. Depending on the chosen model, different prompt endings worked well in making the respective LLM generate valid answers, that is, a number between 1 and 7. After considering the output probabilities over a subset of 15 input samples, we ended up using either prompt ending “Only answer with the score:” or “I choose the number”. In each case, the chosen prompt ending guarantees that more than 99% of the probability mass is allocated to the valid answers. For GPT-4,

Exp.	Image Schema	LLaMA-2		GPT			Avg.
		13b-chat	70b-chat	GPT-3 _{base}	GPT-3 _{inst}	GPT-4	
Gibbs _{stand}	VERTICALITY	0.26	0.41*	N/A	0.53*	0.69*	0.47
	BALANCE	0.27	0.38	-0.05	0.37	0.49*	0.29
	CENTER-PERIPHERY	0.20	0.36	N/A	0.82*	0.56*	0.49
	LINKAGE	0.24	-0.06	N/A	0.46*	0.61*	0.31
	RESISTANCE	0.41*	0.48*	N/A	0.71*	0.82*	0.60
	Avg.	0.28	0.31	N/A	0.58	0.63	
Gibbs _{syn}	VERTICALITY	0.22	0.57*	N/A	0.49*	0.70*	0.49
	BALANCE	0.36	0.50*	0.17	0.50*	0.54*	0.41
	CENTER-PERIPHERY	0.22	0.32	0.16	0.67*	0.67*	0.41
	LINKAGE	0.61*	0.32	N/A	0.61*	0.24	0.45
	RESISTANCE	0.30	0.54*	N/A	0.61*	0.77*	0.56
	Avg.	0.34	0.45	N/A	0.61	0.58	
Beitel _{on}	SUPPORT	0.19	0.32	N/A	0.48*	0.62*	0.40
	PRESSURE	0.37*	0.72*	N/A	0.79*	0.37*	0.56
	CONSTRAINT	0.49*	0.37*	N/A	0.60*	0.47*	0.48
	COVERING	-0.15	0.41*	N/A	0.46*	0.68*	0.35
	VISIBILITY	0.24	0.38*	N/A	0.69*	0.62*	0.48
	Avg.	0.23	0.44	N/A	0.60	0.55	

Table 2: Spearman correlation: model answers and human answers. * for $p < 0.05$, **bold** = highest correlation

the log-probabilities are not accessible through the API, which is why we reused the prompt ending chosen for davinci-003. Based on the 15 tested samples, we also assessed whether the models are very sensitive to these scores in relation to the number they output. Fortunately, we found that the scores are relatively stable and, on average, differ only around half a point on the 7-point scale.

Answer Correlations. First, we compare the two evaluation methods presented in Section 4.3, i.e., simply extracting the most likely answer or computing the average of all valid answers based on their likelihood. Using davinci-003, we find that, on average, the thereby extracted answers are only 0.39 points apart. Thus, we decided to use the simpler method of using the most likely answer when presenting the results from hereinafter since it has the advantage of incurring fewer computational costs.

Table 2 shows the Spearman correlations between human and model answers for the five different image schemas. For 70% of the image schemas,

GPT-4 generates the answers most similar to those of human participants, with GPT-3_{inst} generating the most similar answers for the remaining 30%. On average, GPT-4 shows a correlation of 0.61, which can be interpreted as moderate. In comparison, the LLaMA-2 models produce answers that are more dissimilar to those of humans, as indicated by some moderate correlations and many correlations that are not statistically significant. The 13b variant achieves an average correlation of 0.31, and the 70b variant of 0.38. These lower correlations can be partly attributed to the fact that the LLaMA-2 models tend to answer with either 4 or 7, largely ignoring other options on the scale. The only model that is trained without reinforcement learning from human feedback, GPT-3_{base}, fails to generate insightful answers, nearly always generating a 4.

Despite GPT-4 showing the highest correlations, it still generates some answers strongly deviating from human spatial intuitions. For example, GPT-4 fails to relate a sense of VERTICALITY to the

phrase “the barometer stands at 30 inches” (scoring it 1 compared to the participant average of 4.71) or BALANCE to the phrase “the clock stands on the mantle” (1 compared to 4.46). Generally, it can be observed that many of the outliers are caused when the model gives too low ratings, i.e., 1 or 2.

5.2 Experiment 2 — Beitel et al. (2001)

Results are similar to the previous experiment, with the answers of GPT-4 and GPT-3_{instruct} showing the highest similarity to those of human participants, ranging between correlations of 0.4 and 0.8.

5.3 Experiment 3 — Richardson et al. (2001)

Prompt Selection. In contrast to the previous experiments, the model output is not an ordinal measure (1-7), but a nominal classification. Analogously, we use a subset of input samples along with various different prompt endings to identify that the chosen prompt ending guarantees that a high probability mass is allocated to the labels.

Notable results can be reported for text and vision prompting. For text-based models, the addition of quotation marks around the label (e.g. 'up') increases the probability of the model to choose a valid label if prompted with a quote at the end. In general, using a “Question:” and “Answer:” structure improves label likelihood. For VLMs we cannot obtain the log-probabilities, therefore we follow the examples provided by the model developers³.

Results of Primary Directionality. For all models and all conditions, we summarise the Spearman correlations in Table 3 (detailed results for all choices are listed in the Appendix Tab. A6). We can observe strong correlations (> 0.7) in the textual and pseudo-visual conditions, but not in the visual one. In the textual condition, all models show a significant correlation with the human choices with respect to their choice of a primary directionality (horizontal/vertical). Here, GPT-3_{inst} shows a strong correlation of 0.72. Correlations for pseudo-visual conditions are higher except for one outlier: GPT-3_{base}, which has mostly selected the \uparrow for all items irrespective of the order of choices (i.e. order of arrows in the prompt). On the contrary, GPT-4 has the highest correlation across all models and conditions for the pseudo-visual task with 0.82 (horizontal) and 0.83 (vertical). In the visual condition, both open-source versions of IDEFICS do

³<https://huggingface.co/docs/transformers/main/en/tasks/idefics>

not show any correlation with the human responses. GPT-4_{vision} achieves a significant, but moderate correlation with 0.57 (horizontal) and 0.56 (vertical). On average, correlations between model and human answers are higher in the pseudo-visual condition despite the outlier of the GPT-3_{base} model.

6 Discussion

We explored if LLMs capture human intuitions about image schematic basis in language through three psycholinguistic experiments comparing LLMs and VLMs. Model responses often correlate with people’s, especially in larger models, although discrepancies exist for certain image schemas. Nevertheless, the models reflect spatial primitive intuitions, potentially stemming from their ability to model words, their contextual use, and their relation to schema definitions. Another possibility is that the original experiment papers, serving as training data, might contribute to the models’ reproduction of observed patterns, although parsing the original papers’ results effectively is unlikely.

At the same time, one might wonder why, in some cases, the model answers are so far apart from human answers. Besides the explanation of the lack of embodied experience, one aspect to consider is that the LLMs only had access to one item at a time when rating stimuli. Thus, models are unable to rate items relative to each other — a strategy a human participant is likely to adopt. For instance, given the two stimuli “stand in awe” and “the clock stands on the mantle” separately, GPT-4 gives scores of 2 and 1 for the image schema BALANCE. People, on the other hand, tend to relate BALANCE much more strongly to a clock standing on a mantle. This intuition is captured when prompting the model with both stimuli at once, thus allowing the model to provide relative scores; in this case, the model provides a score of 2 for the first phrase but a score of 4 for the clock on the mantel, thus being much closer to human scores. However, prompting the model to provide all answers at once is currently not feasible, as described in Section 4. This missing comparison between items is also reflected in the fact that the models tend to use the extremes of the scale, i.e., rating items with 1 or 7. Another possible answer for the differences is related to the partially small participant pool used in the original experiments. The three experiments recruited 27, 24, and 173 participants. Therefore, the originally recorded human

TEXT	Llama-2-13b _{chat}	Llama-2-70b _{chat}	GPT-3 _{base}	GPT-3 _{inst}	GPT-4	Avg.
HORIZ. / VERT.	0.53* / 0.54*	0.59* / 0.59*	0.68* / 0.68*	0.72* / 0.72*	0.58* / 0.58*	0.57 / 0.59
PSEUDO-VISUAL	Llama-2-13b _{chat}	Llama-2-70b _{chat}	GPT-3 _{base}	GPT-3 _{inst}	GPT-4	Avg.
HORIZ. / VERT.	0.68* / 0.79*	0.51* / 0.51*	0.17 / 0.16	0.68* / 0.68*	0.82* / 0.83*	0.61 / 0.66
Avg.	0.61 / 0.67	0.55 / 0.55	0.43 / 0.42	0.70 / 0.70	0.70 / 0.71	
VISUAL	IDEFICS-80b _{base}	IDEFICS-80b _{inst}	GPT-4 _{vision}	Avg.		
HORIZ. / VERT.	0.00 / 0.00	-0.01 / -0.01	0.57* / 0.56*	0.19 / 0.18		

Table 3: Spearman correlation: model answers and human answers. * for $p < 0.05$, **bold** = highest correlation

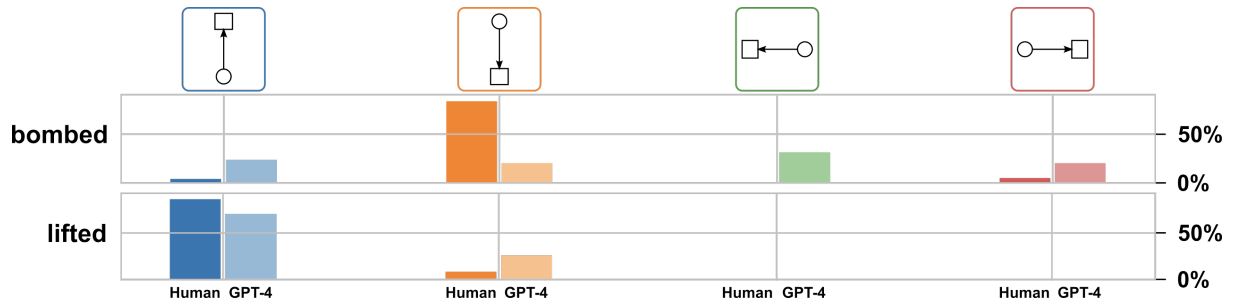


Figure 3: Distribution of image schema choice for items “bombed” and “lifted” by humans (bold) and GPT-4 (light).

answers might not be very robust, and future experiments should rerun similar setups not only with novel stimuli but also with larger participant pools to enable a more robust data set for comparison.

The overall weakest correlations are observed for VLMs, particularly with open-source model IDEFICS, displaying meaningless responses. For instance, IDEFICS-80b_{inst} consistently selects the left-arrow image, disregarding the item, image order, or randomized labels. This failure is attributed to VLMs being trained on natural images, unsuitable for interpreting the highly abstract line drawings. Even GPT-4_{vision}, despite showing moderate correlations (0.56-0.57), occasionally provides random answers, deviating from human consensus (see Fig. 3). This discrepancy raises questions about the models’ alignment with human intuitions.

7 Conclusion

This exploratory study is the first of its kind to reproduce psycholinguistic experiments in order to explore spatial schema intuitions. Moreover, it provides evidence that LLMs are able to reflect those intuitions in different tasks and setups. The results also point out that despite the duo-modality, VLMs do not encode the spatial understanding as effectively as their textual origins. Future research aims to assess models using novel, collected stimuli, ensuring no prior exposure during training. Our cur-

rent findings provide a foundation for formulating precise, testable hypotheses in subsequent experiments. Additionally, we would like to extend this line of research to include a multilingual analysis. In conclusion, our study not only sheds light on the disparities between LLMs and human cognition but also paves the way for new research perspectives in understanding and refining language models. By providing empirical evidence of these disparities, we advocate for a deeper exploration into the limitations of current models and the development of novel approaches to bridge the gap between artificial intelligence and human intelligence.

8 Limitations

While our investigation into the applications of proprietary models, such as GPT-3 and GPT-4, offers valuable insights, it is essential to acknowledge the inherent limitations associated with their use. The opaque nature of the underlying mechanisms in proprietary models poses a significant limitation. The understanding is constrained by the lack of detailed information on the architectural intricacies, leaving us to make assumptions based on analogies to open-source models. It is crucial to recognize that variations in architecture, data, and parameters between proprietary and open-source models impact the generalizability of our findings.

Furthermore, the original experiments have a limited demographic, which we hereby report: (Gibbs et al., 1994) with 27 undergraduate students, U.S. university, native English speakers. (Beitel et al., 2001) with 24 undergraduate students, U.S. university, native English speakers. (Richardson et al., 2001) with 173 undergraduate students, U.S. university, no further information is provided.

Additionally, our reliance on psycholinguistic data introduces notable limitations. Firstly, the temporal aspect of the data is a concern, given its age range of 30 to 23 years. Language evolves over time, and the potential disparities between our data and contemporary linguistic trends may affect the applicability of our results. Therefore, we plan to replicate the studies to gauge temporal robustness. Moreover, we acknowledge that the reproduced studies solely feature the English language and multilingual analysis is subject of future work.

Secondly, the incorporation of original papers into the training data of proprietary models, particularly the LLM and VLM, poses challenges. This integration may introduce biases, potentially influencing the outcomes of our experiments. Yet, the original papers’ results are presented in formats that are unlikely to have undergone parsing during the training procedures, especially within the context of the VLM, yet this remains hard to prove.

9 Considerations and Impact of the Work

Environmental Impact Overall, text generations with the OpenAI API cost 25.85\$. Text generations with LLaMA models were run on two clusters. In those experiments, we utilized 8 NVIDIA RTX A6000 (48GB) GPUs for a 4-hour runtime, each with a power consumption of 300 W, resulting in an estimated total power consumption of 9.6 kWh and a CO2 emission of approximately 3.984 kg. Additionally, we used 2 NVIDIA A100 (40GB) GPUs for a 39-hour runtime, also consuming 300 W each, contributing to a combined total power consumption of 33 kWh and a CO2 emission of ~13.7 kg.

Acknowledgements

Lennart Wachowiak was supported by UK Research and Innovation (EP/S023356/1), in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence, as well as the King’s Institute for Artificial Intelligence.

References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.
- Haldun Akoglu. 2018. User’s guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.
- Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.
- Dinara A Beitel, Raymond W Gibbs Jr, and Paul Sanders. 2001. The embodied approach to the polysemy of the spatial preposition on. In *Polysemy in cognitive linguistics*, pages 241–260. John Benjamins.
- 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, Online. Association for Computational Linguistics.
- Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- Lera Boroditsky. 2000. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28.
- Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR.
- Rodney A Brooks. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.
- Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*. DeepMind.

- Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.
- Ezequiel A Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic bodies: The continuity between life and language*. MIT press.
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT*, pages 32–42.
- Raymond W Gibbs. 2005. The psychological status of image schemas. *From perception to meaning: Image schemas in cognitive linguistics*, 29:113–136.
- Raymond W Gibbs, Dinara A Beitel, Michael Harrington, and Paul E Sanders. 1994. Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of semantics*, 11(4):231–251.
- Dagmar Gromann and Maria M Hedblom. 2017. Kinesthetic mind reader: A method to identify image schemas in natural language. *Proceedings of Advancements in Cognitive Systems*.
- Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2022. Machine intuition: Uncovering human-like intuitive decision-making in gpt-3.5. *arXiv preprint arXiv:2212.05206*.
- Beate Hampe. 2005. Image schemas in cognitive linguistics: Introduction. *From perception to meaning: Image schemas in cognitive linguistics*, 29:1–12.
- Jacqueline Harding, William D’Alessandro, NG Laskowski, and Robert Long. 2023. Ai language models cannot replace human research participants. *AI & SOCIETY*, pages 1–3.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- Conor Houghton, Nina Kazanina, and Priyanka Sukumar. 2023. Beyond the limitations of any imaginable mechanism: large language models and psycholinguistics. *arXiv preprint arXiv:2303.00077*.
- Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. 2023. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint arXiv:2311.09048*.
- Mark Johnson. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago press.
- Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What’s “up” with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.
- Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.
- Jean M Mandler. 1992. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Daniel C Richardson, Michael J Spivey, Shimon Edelman, and Adam J Naples. 2001. " language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.
- Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- C. Spearman. 1904. [The proof and measurement of association between two things](#). *The American Journal of Psychology*, 15(1):72–101.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Lennart Wachowiak and Dagmar Gromann. 2022. [Systematic analysis of image schemas in natural language through explainable multilingual neural language processing](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5571–5581, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. [Drum up SUPPORT: Systematic analysis of image-schematic conceptual metaphors](#). In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.

Philipp Wicke. 2023. Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models. *Findings of the Association for Computational Linguistics: ACL2023*.

Appendix

Image Schema	Definition/Introduction
BALANCE	Consider the notion of BALANCE. Balance refers to your sense of symmetry or stability relative to some point within your body.
VERTICALITY	Consider the notion of VERTICALITY. Verticality refers to the sense of an extension along an up–down orientation.
CENTER–PERIPHERY	Consider the notion of CENTER–PERIPHERY. Center–periphery refers to the experience of some objects or events as central while surrounding objects and events are peripheral or to the outside.
RESISTANCE	Consider the notion of RESISTANCE. Resistance refers to the experience of your body opposing some external force.
LINKAGE	Consider the notion of LINKAGE. Linkage refers to the perception of a connection between objects or events.

Table A1: Image schema definitions provided in the prompts (Experiment 1)

Image Schema	Definition/Introduction
SUPPORT	In the case of the use of “on” in “the book is on the desk”: the SUPPORT relation refers to the desk supporting the book.
PRESSURE	In the case of the use of “on” in “the book is on the desk”: the PRESSURE relation refers to the book exerting some pressure on the desk.
CONSTRAINT	In the case of the use of “on” in “the book is on the desk”: the CONSTRAINT relation refers to the desk constraining the possible motions of the book.
COVERING	In the case of the use of “on” in “the book is on the desk”: the COVERING relation refers to the book concealing the part of the desk that is under the book.
VISIBILITY	In the case of the use of “on” in “the book is on the desk”: the VISIBILITY relation refers to the book being visible on the desk.

Table A2: Image schema definitions provided in the prompts (Experiment 2)

Stimuli with stand (Experiment 1)	Stimuli with synonym (Experiment 1)	Stimuli with on (Experiment 2)
stand at attention	be at attention	The family depends on the father
stand out in several sports	be distinguished in several sports	There is a physician on call
to stand firm	to hold firm	All books are on sale
don't stand for such treatment	don't allow such treatment	The band is on tour
to stand the test of time	to pass the test of time	The boat is on course
united we stand	united we are strong	The bus is on schedule
we stand on 30 years of experience	we are backed up by 30 years experience	Jeff is on time
let the issue stand	let the issue remain as is	Sam is on his way home
let the mixture stand	leave the mixture undisturbed	She puts the blame on my actions
get stood up for a date	have a date with someone who didn't show up	He pulled a gun on me
he stands six-foot nine	he measures six-foot nine	Pat has been on sick leave
the clock stands on the mantle	the clock is on the mantle	There is a parade on Sunday
one-night stand	one-night fling	The program will be broadcast on CBS
to stand to profit	to be in the position to make a profit	Joan works on the committee
to stand in someone else's shoes	to be in someone else's shoes	Linda is very knowledgeable on this subject
on the witness stand	on the witness platform	The factory workers are on strike for a second day
stand in awe	be in aw	These cold nights are very hard on the homeless
the police told them to stand back	the police told them to get back	They ordered a court martial to be held on him
stand by your man	support your man	Pam is on a diet
the engine can't stand the constant wear	the engine can't endure the constant wear	The dog is on the leash
stand on shaky ground	to be on shaky ground	She is on the pill
to stand accused	to be accused	The ship is on the anchor
the house stands in the clearing	the house is in the clearing	The lunch is on George
the barometer stands at 30 inches	the barometer is at 30 inches	He lives on a pension
as the matter now stands	as the matter now exists	The boss is on my neck
the part stands for the whole	the part represents the whole	There are fifty nurses on the hospital staff
it stands to reason	it conforms with reason	There is ten cent interest on the dollar
they did nothing but stand around	they did nothing but hang around	The vase is on the table
to stand against great odds	to face great odds	The fish is on the surface
a standing ovation	a roaring ovation	The fly is on the ceiling
the boss always stands over me	the boss always hovers over me	She has a beautiful ring on her finger
he stands committed	he remains committed	The house rests on the foundation
		They hung a picture on the wall
		I have a cut on my finger
		I have a rash on my back
		The actor is on stage
		He resides on the continent

Table A3: Items the LLMs had to rate in experiments 1 and 2

Action words (Experiment 3)

fled, pointed at, pulled, pushed, walked, hunted, impacted, perched, showed, smashed, bombed, flew, floated, lifted, sank, argued with, gave to, offended, rushed, warned, owned, regretted, rested, tempted, wanted, hoped, increased, obeyed, respected, succeeded

Table A4: List of action words from (Richardson et al., 2001). These words are used for Experiment 3.

Condition	Prompt
TEXTUAL	Given the concepts: [DIRECTION WORDS]. For the concept that best represents the event '[ACTION WORD]', I would choose '
PSEUDO-VISUAL	Given the event '[ACTION WORD]', which of the following arrows best represents this event: [UNICODE ARROWS]. A research participant would choose the arrow:
VISUAL	[Instruction: You are shown four images. Select one of the four images to answer the question.] Image [IMAGE LABEL]:[IMAGE] Image [IMAGE LABEL]:[IMAGE] Image [IMAGE LABEL]:[IMAGE] Image [IMAGE LABEL]:[IMAGE] Question: Which is the best image to describe the event '[ACTION WORD]'? Answer: For '[ACTION WORD]' it is Image
GPT-4 _{CHAT}	[TEXTUAL or PSEUDO-VISUAL PROMPT] You are a participant in a research experiment. Even if the answer is subjective, provide it. Do not say it is subjective.

Table A5: Prompts of the three different modality-conditions used in Experiment 3. DIRECTION WORDS: 'up', 'down', 'left', 'right'. UNICODE ARROWS: ↑, ↓, ←, →. IMAGE LABEL: VMBR, WJZX, XQHL, YGPK. IMAGE see Fig. 1 (without letter labels).

Condition	Image Schema	LLaMA-2		GPT			Avg.
		13b-chat	70b-chat	GPT-3 _{base}	GPT-3 _{inst}	GPT-4	
TEXTUAL	UP	0.49	0.66	0.67	0.63	0.66	0.62
	DOWN	0.21	0.23	0.34	0.31	0.33	0.28
	LEFT	-0.12	0.25	0.23	0.37	0.24	0.24
	RIGHT	0.57	0.56	0.53	0.56	0.41	0.53
PSEUDO-VISUAL	↑	0.44	0.61	-0.09	0.49	0.70	0.43
	↓	0.44	0.14	0.31	0.42	0.49	0.36
	←	0.18	0.29	0.03	0.31	0.18	0.20
	→	0.31	0.43	N/A	0.56	0.69	0.50

Table A6: Spearman correlation between model answers and human answers. **bold** = highest correlation

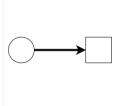
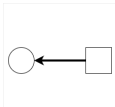
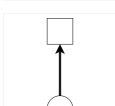
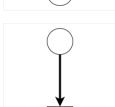
	Image	IDEFICS		GPT	Avg.
		80b	80b-inst	GPT-4	
VISUAL		0.12	0.10	0.51	0.24
		-0.12	0.01	0.27	0.05
		-0.23	0.02	0.11	-0.03
		0.49	0.30	0.53	0.44

Table A7: Spearman correlation between model answers and human answers. **bold** = highest correlation