



PDF Download
3778353.pdf
18 December 2025
Total Citations: 0
Total Downloads: 51

 Latest updates: <https://dl.acm.org/doi/10.1145/3778353>

RESEARCH-ARTICLE

Sensorimotor Regularities as Alignment between Humans and Large Language Models

JINGYI LI, University of Cambridge, Cambridge, Cambridgeshire, U.K.

JINGHUI HU, Lancaster University, Lancaster, Lancashire, U.K.

PER OLA KRISTENSSON, University of Cambridge, Cambridge, Cambridgeshire, U.K.

Open Access Support provided by:

Lancaster University

University of Cambridge

Accepted: 02 November 2025

Revised: 30 October 2025

Received: 08 October 2024

[Citation in BibTeX format](#)

Sensorimotor Regularities as Alignment between Humans and Large Language Models

JINGYI LI, University of Cambridge, United Kingdom

JINGHUI HU, Lancaster University, United Kingdom

PER OLA KRISTENSSON, University of Cambridge, United Kingdom

Large Language Models (LLMs) do not construct conceptual representations in ways that align with human cognition, posing risks for human-AI interaction. While LLMs solely rely on linguistic distributional knowledge, humans leverage both linguistic and sensorimotor knowledge. To systematically assess human-LLM alignment in concept representations, we propose a novel evaluation framework based on sensorimotor regularities, operationalized as image schemas—multimodal gestalts derived from repeated sensorimotor experiences. Investigating linguistic manifestations of such schemas, we systematically identify human-LLM alignments and misalignments in the encoding of sensorimotor regularities. Results indicate that three contemporary disembodied LLMs encode highly human-like sensorimotor gestalts. However, these models exhibit reduced alignment when mapping such gestalts to concepts, and they do not systematically combine these gestalts in ways consistent with human patterns. We identify each LLM’s misalignments with human patterns in image schema distribution, conceptual associations, and image schema co-occurrences. Building on these findings, we augment gpt-4-1106 with targeted sensorimotor priors derived from its identified misalignments with human patterns. In a downstream user study, this augmentation yields sentence continuations rated by humans as significantly more conceptually clear, contextually contingent, and human-like than baseline outputs. Our work establishes a foundation for evaluating and improving human-LLM alignment at the conceptual level.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Artificial intelligence**; **Knowledge representation and reasoning**;

Additional Key Words and Phrases: Large Language Models, Alignment, Concept Representation, Image Schema, Cognitive Linguistics

1 Introduction

Large language Models (LLMs) have impressed the world with their capabilities in generating human-like language. This naturally leads to the question: to what extent can LLMs approximate patterns of human conceptual reasoning? One way to understand this is to examine the similarities and differences in how LLMs and humans represent *concepts*, which refer to mental representations that allow the categorization of a group of nonidentical but equivalent entities [28]. Concepts are considered the foundation of the human cognitive system [14], enabling us to categorize and structure our understanding of the world, to infer similarities based on category membership, and to make predictions about objects and events [5, 34].

LLMs and humans employ distinct strategies for representing concepts. Leveraging linguistic distributional knowledge, LLMs represent a concept based on its relationship to other words [15]. However, human conceptual

Authors’ Contact Information: Jingyi Li, University of Cambridge, Cambridge, United Kingdom, jl2231@cam.ac.uk; Jinghui Hu, Lancaster University, Lancaster, United Kingdom, j.hu23@lancaster.ac.uk; Per Ola Kristensson, University of Cambridge, Cambridge, United Kingdom, pok21@cam.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, or post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s).

ACM 1557-7325/2025/11-ART

<https://doi.org/10.1145/3778353>

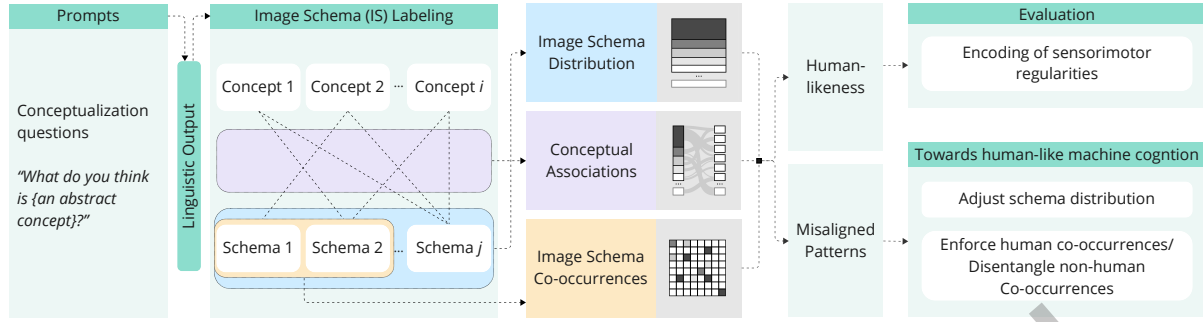


Fig. 1. The framework for the evaluation of sensorimotor regularities encoded by LLMs.

representation relies on both linguistic distributional knowledge and sensorimotor knowledge from the physical world [7, 14, 51]. Similar to the statistical regularities in language, humans develop a set of regularities abstracted from reoccurring sensorimotor experiences starting in early infancy [53–55]. These sensorimotor regularities are known as *image schemas* [6, 40, 44]. We use them as conceptual building blocks to understand unfamiliar concepts that lack physical referents. For example, the image schema *CENTER-PERIPHERY* is abstracted from the bodily relations between our trunk and extremities. This schema helps us understand abstract concepts not grounded in the physical world, such as the notion of *Importance*. The recurring experiences that central parts of the body are vital, while peripheral parts, like hair, which can be easily removed without harm, are less important, leads us to associate *CENTER* with *Importance*, and *PERIPHERY* with *Unimportance*. Such associations are instantiated and reinforced in human behaviors (e.g., centering key messages in advertisements) and language (e.g., “Let’s focus on the *CENTRAL* issue and set the rest *ASIDE*”).

Unlike humans, who can leverage both linguistic distributional knowledge (e.g., the concept of *Learning* could activate related words like *memorization*) and sensorimotor knowledge (e.g., *Learning* could activate image schema *PATH*¹) to represent abstract concepts, LLMs exhibit proficiency in linguistic distributional knowledge but inherently lack grounding in the physical world. A crucial aspect of human-LLM collaboration is alignment, with most discussion on LLM alignment focusing on aligning LLMs to human values [12, 22, 35]. However, researchers contend that before attempting to align values, it is essential to align the concept representations encoded in LLM outputs with those in human understanding [60]. To evaluate the concept alignment between LLMs and humans, it is important to answer this open question: to what extent might LLMs encode the sensorimotor regularities that humans leverage in concept representation?

Leveraging the manifestations of image schemas in language, we operationalize the *sensorimotor regularities* in the above inquiry into three quantifiable metrics that can be elicited from the language of any linguistic agent, artificial or biological: (1) the global distribution of image schemas; (2) conceptual associations, which we define as the association made between an abstract concept and an image schema; and (3) the co-occurrences between any image schema pairs under the same concept. These metrics can be understood via a toolbox analogy: the first metric represents the variety and quantity of tools available, the second indicates how different tools are used, and the third shows which tools are usually used together.

This paper develops an evaluation framework (see Fig. 1) that enables the quantification of these metrics for any linguistic agent, allowing for statistical comparisons between an LLM and humans. This framework can reveal

¹*Learning* is often conceptualized as having a starting point and thereafter moving through a sequence of steps toward a final goal. This aligns structurally with the *PATH* schema, which involves a source, a destination, and a sequence of contiguous locations connecting the source with the goal [40]. Linguistic expressions such as “on the path to understanding,” “step-by-step learning,” or “progress through the material” reflect this underlying conceptual association.

the humanlikeness of the sensorimotor regularities encoded by a model and specify any deviations from typical human patterns. This framework is applied on four linguistic agent types (LLaMa-2-7b [64], LLaMa-2-70b [64], gpt-4-1106 [56], and humans ($n=24$)) by zero-shot prompting them with the same set of conceptualization questions in the domain of technology learning, and then labeling image schemas from their linguistic responses.

The alignment evaluation results suggest that all three models encode largely human-like image schema distributions, with a small number of image schemas (e.g., *ENABLEMENT* (for a definition, see Table 9 in the Appendix), *COLLECTION*, and *SURFACE*) being disproportionately used. All three models demonstrated decreased humanlikeness when making conceptual associations between image schemas and concepts, but each model produced at least one human-like association for every concept. Further, the performance of all three models decreases when capturing the complex interactions between image schema pairs embedded in human cognition. For example, when humans use *ENABLEMENT* to understand a concept, they tend to also use *IN*. The co-occurrence between *ENABLEMENT-IN*, however, was not replicated by any of the three models. In section 5, we report detailed insights on the similarities and differences between the sensorimotor regularities encoded by each of the three models and those embedded in human language.

Drawing on findings from our alignment evaluation, we further investigated whether closer sensorimotor alignment could translate into practical benefits in downstream applications. Through a creative writing task with fictional terms, we demonstrate that gpt-4-1106—when guided by targeted sensorimotor priors derived from its specific misalignments with human patterns—produces sentence continuations that human evaluators ($n=45$) rate as conceptually clearer, more contextually contingent, and more human-like, compared to a baseline gpt-4-1106 without access to these additional priors. This proof-of-concept demonstration provides an initial signal of the downstream benefits that can be achieved by bringing LLMs to closer alignment with humans through sensorimotor regularities.

We note that although our analyses compare LLM outputs with human conceptualizations, we do not assume LLMs possess understanding or embodied cognition. In line with prior work, we regard LLMs as statistical systems that generate plausible text without semantic comprehension or cognitive grounding [8, 9], and interpret any resemblance between human language and LLM output as purely functional rather than indicative of shared cognitive processes [16]. Our goal is to assess the functional resemblance between LLMs and humans in specific aspects of conceptual representation, without anthropomorphizing the models.

In summary, this paper makes the following contributions:

- A systematic evaluation framework utilizing image schemas to assess how well a language model without direct grounding encodes sensorimotor regularities in human cognition, and more importantly, identifies specific misalignments.
- An empirical alignment analysis of three LLMs (LLaMa-2-7b, LLaMa-2-70b, gpt-4-1106) using this framework, revealing that while all three demonstrate considerable capability in encoding surface-level characteristics (e.g., image schema distribution) of human sensorimotor regularities in their outputs, they exhibit limited capacity to capture more intricate internal interactions and dynamics between these sensorimotor constructs. Such capabilities and limitations can offer specific directions for better aligning LLMs to human cognition and facilitating their cross-modal reasoning capabilities.
- A proof-of-concept downstream study demonstrating that augmenting LLMs with targeted sensorimotor knowledge derived from its empirically identified misalignments with human patterns leads to outputs that human evaluators rate as conceptually clearer, more contextually contingent, and more human-like—demonstrating the practical value of sensorimotor alignment in real-world tasks.

2 Image Schema Theory

2.1 Bridging the Body, Thought, and Language

Rooted in cognitive linguistics and embodied cognition research, image schemas are defined as “recurring dynamic pattern(s) of our perceptual interactions and motor programs that gives coherence and structure to our experience” [40]. The psychological reality of image schemas is supported by an abundance of evidence from research in psycholinguistics [11, 24, 26, 61] and developmental psychology [3, 4, 20, 21, 45, 46, 49, 53–55, 69]. Prior research has identified 30–40 image schemas, and in later work, Hurtienne and Blessing [36] grouped them into seven categories based on similarity (see Table 8 in the Appendix).

Starting in early infancy, image schemas act as the fundamental thread that weaves together our sensorimotor experiences, thought, and language [24]. Human infants come equipped with a mechanism known as *Perceptual Meaning Analysis*, which analyzes perceptual information and recodes them into image schemas, a vastly simpler and reduced form [55]. For example, infants have the experience of repeatedly putting items inside another object (e.g., a box) and taking the items out, or simply observing others perform these actions. Through perceptual meaning analysis, these experiences are abstracted into an image schema CONTAINER that consists of an interior, an exterior, and a boundary [18]. Such image schemas represent a set of sensorimotor regularities in our experiences—they possess inherent structure, and can be metaphorically extended to provide understanding for other abstract concepts [25, 40, 43]. The CONTAINER schema, for instance, give rise to our metaphorical structuring of an abstract concept *Mind*, which subsequently constrains our use and understanding of language with regards to this concept (e.g., “*I can’t get this idea out of my mind.*”) [23].

Serving as the bridge between body–thought–language, image schemas provide a promising approach for reverse engineering insights into bodily experiences and cognition from language. Both humans, who possess physical bodies, and language models, which lack physical form, can use language to discuss certain concepts. Through analyzing the linguistic manifestations of image schemas, we can compare the embodied structures inferred from human language use with those statistically embedded in language model outputs. Investigating what kind of embodied structures a language model implicitly approximates—relative to the human case—may offer valuable insight for improving model performance in human-designed tasks that require embodied understanding and interaction.

2.2 Quantifiable Metrics

We propose to evaluate the humanlikeness of sensorimotor regularities encoded by language models and deviations from human patterns across three quantifiable, progressively sophisticated metrics.

Starting with *image schema distribution*, a metric that examines the presence and frequency of each image schema, we set the groundwork by identifying the conceptual building blocks available. The second metric, *conceptual associations* between image schemas and various concepts, takes one step further by examining how these available building blocks are re-activated to structure the representations of abstract concepts. Finally, *image schema co-occurrences* represent the most sophisticated patterns that emerge from repeated re-activation of different image schemas during different conceptualization events. Here we further define the metric of image schema co-occurrences and discuss how it can facilitate cross-modal and cross-domain reasoning.

One image schema can often co-occur with another image schema under the representation of the same abstract concept. For example, image schemas such as BIG and HEAVY are conceptually associated with the same abstract concept *Importance* (e.g., “*Tomorrow is a BIG day*”, “*The findings add WEIGHT to our research*”). Interestingly, BIG and HEAVY are usually used together to represent the properties of another abstract concept *Responsibility* (e.g., “*Leading this project is a MASSIVE task*”, “*She carries the WEIGHT of the world on her shoulders*”). Humans experience in the physical world that BIGGER things usually (although not always) tend to be HEAVIER. The activation of

BIG during abstract conceptualization can very likely also invoke HEAVY, and vice versa, leading to a frequent co-occurrence between this image schema pair in the representations of various concepts.

The intuition of schema co-occurrences emerges from recurrent physical experiences and repeated conceptualization events, deeply ingrained in human cognition. If correctly encoded, this intuition could improve a language model’s capacity to extend patterns across domains beyond language, allowing it to infer plausible sensorimotor properties of unseen concepts based on known image schemas. For instance, by recognizing patterns such as the frequent co-occurrence of ENABLEMENT and IN, a model can infer possible spatial relationships from language alone. Knowing that “entity *A* ENABLES entity *B* to function”, a model could spatially infer that entity *A* could possibly reside INSIDE of entity *B*. It is interesting to investigate to what extent these schema co-occurrences are correctly encoded by language models without direct grounding in the physical world.

Taken together, these three metrics build upon one another, providing a comprehensive understanding on how sensorimotor regularities are encoded in language models.

3 Related Work

We situate our work within two bodies of literature, one that investigates different types world knowledge encoded by language models, and another called “machine psychology” that seeks to interpret LLM behaviors by treating them as human participants in psychological experiments. Below we describe each body of literature.

3.1 World Knowledge Encoded by LLMs

Online comprehension of language involves both linguistic and world knowledge [59]. World knowledge refers to non-linguistic, factual and commonsense knowledge about agents, objects, properties, actions, events, and ideas in the world [52], which enables us to access the plausibility of a linguistic expression in relation to our mental representation of the world’s state of affairs [32, 59]. Efficient and safe human-AI collaboration relies on the accurate encoding of world knowledge in LLMs, which motivates research investigating different types of world knowledge encoded by LLMs. Various studies have shown that representations learned by language models encode factual and commonsense knowledge [17, 57]. Similar efforts investigate how language models encode properties and affordances of objects [19], as well as human stereotypic tacit assumptions [67]. Additionally, prior research explores how language models encode relational knowledge about what entities are related and how they are related [39], as well as human lexical similarity, relatedness judgment, and analogy relationships [13]. Particularly relevant to our work is recent work [68] exploring whether pre-trained language models’ concept representations encode certain spatial schemas (e.g., UP-DOWN), a subset of image schemas. Our work goes beyond examining the correct encoding of a subset of image schemas in language models. Instead, we propose a framework to systematically evaluate the humanlikeness of a language model’s sensorimotor regularities encoding, in terms of image schemas distribution, conceptual associations, and image schema co-occurrences. Developing new methods that enable researchers to monitor the humanlikeness of the conceptual world constructed from LLMs’ language will become increasingly in the context where AI-generated data may continue to contaminate the training data of newer generations of LLMs, a situation demonstrated to cause models to rapidly collapse, producing nonsensical outputs [62].

3.2 Machine Psychology Studies

Driven by the needs to investigate the functional similarities and differences between LLM outputs and human outputs, a growing body of research has begun to compare LLM outputs and human responses on psychological experimental paradigms. We focus on prior research that compares the statistical patterns of LLM outputs with human patterns using psychological experiments, for a comprehensive review of machine psychology studies, see [30]. Prior study tested GPT-3 on a set of classic psychology experiments (e.g. the Linda problem), and

demonstrated human-like errors and content effects in model outputs [10]. Similarly, another study evaluated LLM responses to a battery of semantic illusions and cognitive reflection tests, and observed response patterns resembling human-like system-1 thinking and associated cognitive errors [31]. Moreover, researchers assessed GPT-3.5 and GPT-4 on a classic inductive reasoning task, finding that GPT-3.5 is outperformed by humans while GPT-4's performance matches that of humans [33]. Other traits such as creativity and intelligence have also been investigated. Researchers assessed GPT-3's outputs on the Alternative Uses Task and reported human-level performance in creative ideation [63]. In another study, a text-based matrix reasoning task was given to GPT-3 to measure intelligence, and results indicate that it performs as well or better than humans in most conditions [66]. While prior studies in machine psychology mostly focus on LLMs' ability to replicate specific human cognitive processes like inductive reasoning, we compare LLMs and humans at a more fundamental level, examining how they represent different concepts that define our world via sensorimotor regularities. Since such concepts underpin the human cognitive system, we can infer similarities and make predictions about objects and events. Examining concept alignment between LLMs and humans may help explain where and why model outputs converge with, or diverge from, human cognitive patterns.

4 Methods

Here we describe how our proposed framework can be used for evaluating the humanlikeness of sensorimotor regularities encoded in language models.

4.1 Data Collection

4.1.1 Human Conceptual Representations. A human baseline can be created by collecting conceptual representations of a specific set of abstract concepts through one-on-one interviews with a group of human participants.

In the current study, we reused 24 interviews from a previous study to establish the human baseline. In the prior study [47], we recruited 24 English-speaking participants between the age of 20 and 75 to collect their conceptual representations. The sample consisted of two equal-sized age groups: 12 younger participants aged 21–34 (Mean = 25.92, SD = 3.55), recruited via a mailing list distributed to members of a large university in the UK, and 12 older participants aged 62–75 (Mean = 70.21, SD = 4.04), recruited via a mailing list distributed to members of U3AC, a charity that organizes educational activities for people who are not or no longer in full-time employment. We collected their concept representations through one-on-one interviews, where they were required to describe a set of 20 abstract concepts in their own words, by answering a standardized question, “What do you think is [an *Abstract concept*]?” All participants were instructed to answer the question in the format of “[*Abstract concept*] is *something*”, and to avoid oversimplified answers like “[*Abstract concept*] is *its synonym*”.

The 20 abstract concepts described by participants are *Functionality, Software, Button, Information, Problem, Operation, Barrier, Manual, Screen, Interface, Video Tutorial, System, Option, Search, Learning, Control Panel, Step, Logic, Information Hierarchy, Assistance*. They are a set of high-frequency words identified from a transcribed interview corpus related to technology use and learning [47]. For each of the 24 human participant, we collected 20 concept descriptions, each describing one of the above abstract concepts.

We chose the domain of technology learning as the elicitation context because it encompasses a wide range of abstract concepts that are particularly likely to invite image-schematic structuring. Unlike concrete concepts such as “banana” or “desk”, which tend to evoke direct perceptual representations, abstract notions like *information, logic, or system* encourage participants to recruit metaphorical and embodied reasoning. This makes them well-suited for eliciting image schema-based conceptualizations. While our empirical focus is on this domain, the framework and methodology are designed to generalize beyond it. Future work can readily adapt our approach to study sensorimotor alignment in other abstract conceptual domains.

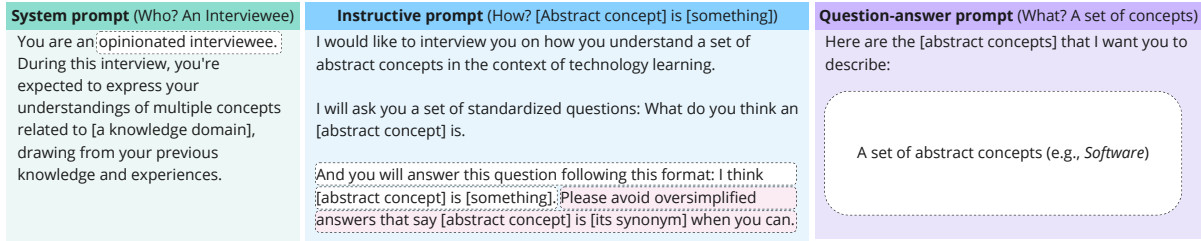


Fig. 2. Prompt structure.

4.1.2 Model Conceptual Representations. Through zero-shot prompting, we elicited responses from three large language models (LLMs) of varying sizes using the same interview questions administered to human participants: LLaMa-2-7b (7 billion parameters), LLaMa-2-70b (70 billion parameters), and gpt-4-1106 (the exact number of parameters is not publicly disclosed, but it is suggested to be bigger than GPT-3.5's 175 billion parameters) [56]. These models were selected to cover a range of parameter scales and to represent two of the most prominent LLM families available at the time of the study. gpt-4-1106 was widely regarded as state-of-the-art for general-purpose reasoning and performance when the study was conducted, while Llama-2 was the leading open-source model family in terms of performance and adoption during the same period.

Prompt Structure. Prompting allows us to present models with the same inputs that are presented to human participants in the interviews. As shown in Fig. 2, we use a *system prompt* to define the intended communicative role, a *question-answer prompt* to specify the set of concepts to be described, and an *instructive prompt* to guide the model's output toward a standardized response format (“[*Abstract Concept*] is *something*”) and to avoid oversimplified answers (“[*Abstract Concept*] is *its synonym*”).

Iterative Prompting Strategy. We evaluated LLM outputs under their default sampling parameters, those that typical users would encounter when interacting with models via public interfaces. All prompts were submitted using the models' default sampling parameters: LLaMA-2-7B and LLaMA-2-70B were queried with temperature = 0.75 and top- p = 0.9, while gpt-4-1106 was queried with temperature = 1.0 and top- p = 1.0.

Under these default settings, LLMs can generate diverse outputs in response to the same prompt due to their probabilistic nature. This variability allows each sampled output to be treated as a distinct *model sample*, analogous to a simulated participant. To manage this diversity, we employed an iterative prompting strategy. Initially, the model was zero-prompted with the above prompt, and a human annotator identified image schemas in each of the 20 concept descriptions within this model's output. This output was treated as one model sample. We then re-prompted the LLM with the same prompt and annotated the new output in the same manner, treating it as a separate model sample. This process was repeated until we reached saturation [27], the point at which additional outputs no longer yielded new image schemas for each concept. Although saturation was defined based on the emergence of new image schemas per concept, this criterion implicitly constrains all three metrics: image schema distributions, schema-concept associations, and schema-schema co-occurrences. Once no new image schemas are generated for a given concept, no new conceptual associations for this concept or schema co-occurrences within this concept can arise. This iterative method enabled comprehensive and representative sampling of the image schemas embedded in the model's conceptual representations. We applied this process to each of the three LLMs, collecting twelve unique outputs per model—yielding twelve distinct model samples for each LLM.

4.2 Image Schema Elicitation

4.2.1 Annotation Process. To manually annotate image schemas in human participants' and models' concept descriptions, we use the ISCAT database [37] as a dictionary, which contains image schemas' definitions collected from the literature and their linguistic instantiations. To support consistent identification, we compiled a list of signaling words for each image schema based on the ISCAT database (see Table 9 in the Appendix). Both the primary annotator (the first author) and the second annotator (the second author) referred to the same annotation guideline during their independent annotation. The second annotator, who has prior experience in image schema elicitation from linguistic data, received a tutorial on schema definitions and their signaling words, as well as step-by-step instructions for identifying image schemas in concept descriptions.

In each concept description (e.g., “*An interface is the link between human beings and the things that they need to operate in the world.*”), the annotators identify image schema(s) based on schema definitions and their associated signaling words. For example, in the provided description of the concept *Interface*, the image schema LINKAGE, which is defined as a linking device connecting two or more entities with each other, was indicated by the word *link*. Other signaling words for LINKAGE also include (but are not limited to) *connect*, *bridge*, and *bond*.

To better capture the collective use of image schemas in conceptualization rather than an individual human participant's preference, we aggregate repeated references to the same image schema by a single human participant when describing the same concept into a single image schema instance. The same approach is applied to an LLM' concept descriptions. If a model sample (i.e., a single response to the interview prompt) contains multiple occurrences of the same image schema in descriptions of a given concept, these are aggregated into a single image schema instance.

4.2.2 Inter-Annotator Agreement. Given the substantial time and effort required for having more than one annotator process the entire dataset, Armstrong et al. [1] recommend that inter-annotator agreement can be reliably evaluated using a representative subset comprising 25% of the entire dataset. Our image schema elicitation study involved a primary annotator independently annotating the entire dataset, with a second annotator later independently annotating a representative 25% subset to assess inter-annotator agreement.

To ensure the subset was representative, we performed random sampling within the data of each participant group (Humans, llama-2-7b, llama-2-70b, gpt-4-1106), sampling 25% of the instances coded with each of the six image schema categories that made up the entire dataset. For instance, within the human data, 25% of instances annotated with image schemas in the SPACE category were sampled. This procedure was applied consistently across all schema categories within the data of each participant group, resulting in a total of 360 instances for the double coding subset.

The second annotator then independently annotated this representative subset. Inter-annotator reliability on this subset was assessed using Krippendorff's alpha, yielding a value of 0.833 (no missing value; 360 decisions; observed disagreement: 0.147; expected disagreement: 0.880), which reflects a substantial level of inter-annotator agreement. We used this agreement value to evaluate the consistency, validity, and reproducibility of the primary annotator's work against the established annotation guideline. This substantial agreement provided assurance that the primary annotator's annotation could serve as a reliable basis for analyzing the entire dataset.

4.2.3 Elicitation Results. As shown in Table 1, from the 480 concept descriptions we collected from 24 human participants, we identified 303 image schema instances, among which 44 instances were only used by only one human participant. From the 240 concept descriptions we collected from 12 distinct model samples from the Llama-2-7b family, we identified 376 image schema instances, with 8 instances appearing in only one model sample. For Llama-2-70b, 240 concept descriptions generated by 12 model samples yielded 346 image schema instances, with 3 instances appearing in only one model sample. Finally, the 240 descriptions generated by 12

Table 1. Number of image schemas (ISs) elicited from human data and three contemporary LLMs.

	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
Concept descriptions	480	240	240	240
All ISs elicited	303	376	346	419
ISs used by only one participant	44	8	3	7

model samples from gpt-4-1106 produced 419 image schema instances, with 7 appearing in only one model sample.

The number of image schema instances identified in the LLM outputs was higher than in the human descriptions. Not every human concept description contained an image schema instance, often due to individual expression habits or brevity. In contrast, when answering the same interview questions, LLMs typically produced more elaborate responses, which naturally introduced more opportunities for schema activation. For example, gpt-4-1106 described *System* as “*I think a system in technology refers to a set of interconnected components that form an organized framework, working together to perform functions or solve problems. Understanding a system’s design and operation is critical for effectively engaging with technology-powered learning environments.*” Some human participants, by comparison, gave very brief answers (e.g., “*It (system) is basically a set of standard instructions.*”, “*This (system) is the background of most computers.*”), which helps explain the higher number of image schema instances in the LLM outputs.

Even though not every human description included an identifiable image schema, we consider image schema patterns a valid lens for establishing a human benchmark of conceptual structures. Our rationale is not that image schemas are necessarily high-frequency linguistic phenomena in all utterances, but rather that they are theoretically robust, widely shared across individuals, and underpin typical human reasoning about abstract concepts and experience. Accordingly, establishing a reliable schema-based benchmark requires identifying patterns that are robust and stable across different demographics. To validate this, we conducted a human–human agreement analysis comparing two demographically distinct subgroups in our sample. The high consistency observed across all three metrics (image schema distribution, conceptual associations, and schema co-occurrences) indicates that these sensorimotor patterns are stable across participant backgrounds. We believe this provides a strong foundation for using schema-based patterns as a reference point for evaluating humanlikeness in LLM conceptualizations.

4.2.4 Data Filtering for Representativeness. To ensure that the extracted patterns from human or model data reflect the group’s consensus, we applied a filtering procedure that excludes image schemas assigned to a concept by only one human participant or one model sample, as suggested by prior research in image schema elicitation [50]. This decision ensures that the schema patterns used in subsequent analyses reflect shared or recurrent conceptual structures. As a result, this filtering may exclude certain concept associations or schema co-occurrences that appear conceptually plausible but did not occur frequently enough to meet the inclusion threshold. This trade-off between representativeness and coverage is a design choice to support robust and interpretable cross-group comparisons.

4.3 Data Analysis

4.3.1 Image Schema Distribution. Human labeling of image schemas in a linguistic agent’s concept descriptions enables us to map out how frequently certain image schemas appear within the data, thereby forming an image schema distribution.

Humanlikeness. We transform the image schema distribution of a linguistic agent, which consists of multiple schema types and their occurrence, into vector form. Each vector component corresponds to the count of occurrence (C_{schema_i}) of a specific image schema.

$$\mathbf{v}_{\text{schemas,agent}} = \begin{pmatrix} C_{\text{schema}_1,\text{agent}} \\ C_{\text{schema}_2,\text{agent}} \\ C_{\text{schema}_3,\text{agent}} \\ \vdots \\ C_{\text{schema}_n,\text{agent}} \end{pmatrix} \quad (1)$$

This allows us to compare two vectors using cosine similarity, which measures the cosine of the angle between the two vectors. The cosine similarity between $\mathbf{v}_{\text{schemas,model}}$ and $\mathbf{v}_{\text{schemas,human}}$ is given by:

$$\cos(\theta) = \frac{\sum_{i=1}^N C_{\text{schema}_i,\text{model}} \times C_{\text{schema}_i,\text{human}}}{\sqrt{\sum_{i=1}^N (C_{\text{schema}_i,\text{model}})^2} \times \sqrt{\sum_{i=1}^N (C_{\text{schema}_i,\text{human}})^2}} \quad (2)$$

Traditionally, cosine similarity yields a value between -1 and 1, where 1 indicates perfect similarity and -1 suggests complete dissimilarity. However, in our analysis, this metric ranges from 0 to 1, since all vector components are non-negative. Using cosine similarity, we quantify the similarity between human schema distribution and that of each model.

In addition to providing a global measure of humanlikeness, we further investigate how a model's behavioral patterns contribute to its success or failure of mimicking human patterns. For example, a more conservative model that replicates only a limited range of human schemas and uses very few unique schemas might achieve higher cosine similarity to humans than a bolder model that replicates a broader range of human schemas while also inventing a significant proportion of unique schemas not found in human data.

We investigate such patterns by analyzing the proportions of image schemas unique to humans ($P_{\text{unique,human}}$), unique to the model ($P_{\text{unique,model}}$), and shared between both ($P_{\text{shared,human}}$, $P_{\text{shared,model}}$).

$$P_{\text{unique,agent}} = \frac{\sum_{i=1}^N C_{\text{unique}_i,\text{agent}}}{\sum_{i=1}^N C_{\text{schema}_i,\text{agent}}} \quad (3)$$

$$P_{\text{shared,agent}} = 1 - P_{\text{unique,agent}} \quad (4)$$

Deviations. To identify disproportionately used image schemas, we evaluate the level of deviation as the proportional difference of a specific schema in two distributions (humans and LLM). The level of deviation for a specific image schema i is calculated as:

$$\Delta_{\text{schema}_i} = |P_{\text{schema}_i,\text{model}} - P_{\text{schema}_i,\text{human}}| \quad (5)$$

where $P_{\text{schema}_i,\text{model}}$ is the occurrence count of schema i divided by the total occurrence counts of all schemas in an LLM's language, $P_{\text{schema}_i,\text{human}}$ is the occurrence count of schema i divided by the total occurrence counts of all schemas in human language.

4.3.2 Conceptual Associations. A conceptual association (*Concept-IMAGE SCHEMA*) can be identified when a linguistic agent represents a *Concept* in terms of an *IMAGE SCHEMA*. Human labeling of image schemas in the linguistic description of each concept enables us to identify which and how frequently certain conceptual associations appear within the linguistic data.

Humanlikeness. Similar to how we compare two distributions of image schemas, we transform conceptual associations made by a linguistic agent, which consists of multiple associations types and their occurrence, into vector form. Each vector component corresponds to the count of occurrence ($C_{\text{association}_i}$) of a specific association type.

$$\mathbf{v}_{\text{associations,agent}} = \begin{pmatrix} C_{\text{association}_1,\text{agent}} \\ C_{\text{association}_2,\text{agent}} \\ C_{\text{association}_3,\text{agent}} \\ \vdots \\ C_{\text{association}_n,\text{agent}} \end{pmatrix} \quad (6)$$

This allows us to evaluate the similarity between $\mathbf{v}_{\text{associations,model}}$ and $\mathbf{v}_{\text{associations,human}}$ using cosine similarity.

Apart from cosine similarity as a global measure, we further determine the proportion of conceptual associations unique to humans ($P_{\text{unique,human}}$), unique to the model ($P_{\text{unique,model}}$), and shared between both ($P_{\text{shared,human}}$, $P_{\text{shared,model}}$).

$$P_{\text{unique,agent}} = \frac{\sum_{i=1}^N C_{\text{unique}_i,\text{agent}}}{\sum_{i=1}^N C_{\text{association}_i,\text{agent}}} \quad (7)$$

$$P_{\text{shared,agent}} = 1 - P_{\text{association}_{\text{unique},\text{agent}}} \quad (8)$$

Deviations. To identify conceptual associations that are prevalent in an LLM's language but not in human language, and vice versa, we calculate the level of deviation for each conceptual association i as follow:

$$\Delta_{\text{association}_i} = |P_{\text{association}_i,\text{model}} - P_{\text{association}_i,\text{human}}| \quad (9)$$

where $P_{\text{association}_i,\text{model}}$ is the occurrence count of association i divided by the total occurrence counts of all associations in an LLM's language, $P_{\text{association}_i,\text{human}}$ is the occurrence count of association i divided by the total occurrence counts of all associations in human language. This allows us to identify the disproportionately made conceptual associations by a model with the greatest extent of deviation from human patterns.

4.3.3 Image Schema Co-occurrence. Human labeling of different image schemas in the same concept description enables us to identify how frequently each pair of image schemas co-occur within the linguistic data. The co-occurrence value of schema_a - schema_b can be calculated by multiplying the occurrence counts of schema_a in the description of a concept, indexed by i , with the occurrences of schema_b in the same concept ($C_{\text{schema}_a}^i \times C_{\text{schema}_b}^i$). To calculate the total co-occurrence value of schema_a and schema_b , the individual co-occurrence values are summed across all concepts N .

$$\text{Cooccurrence}_{a,b} = \sum_{i=1}^N (C_{\text{schema}_a}^i \times C_{\text{schema}_b}^i) \quad (10)$$

Humanlikeness. We transform schema co-occurrences found in a linguistic agent's conceptualization, which consists of multiple image schema pairs and their co-occurrences value, into vector form. Each vector component corresponds to the co-occurrences value (Cooccurrence_i) of a specific image schema pair.

$$\mathbf{v}_{\text{cooccurrences,agent}} = \begin{pmatrix} \text{Cooccurrence}_{1,\text{agent}} \\ \text{Cooccurrence}_{2,\text{agent}} \\ \text{Cooccurrence}_{3,\text{agent}} \\ \vdots \\ \text{Cooccurrence}_{i,\text{agent}} \end{pmatrix} \quad (11)$$

This allows us to evaluate the similarity between $\mathbf{v}_{\text{cooccurrences,agent}}$ and $\mathbf{v}_{\text{cooccurrences,human}}$ using cosine similarity.

Apart from cosine similarity as a global measure, we further determine the proportion of schema co-occurrences unique to humans ($P_{\text{unique,human}}$), unique to the model ($P_{\text{unique,model}}$), and shared between both ($P_{\text{shared,human}}$, $P_{\text{shared,model}}$).

$$P_{\text{unique,agent}} = \frac{\sum_{i=1}^N \text{Cooccurrence}_{\text{unique},i,\text{agent}}}{\sum_{i=1}^N \text{Cooccurrence}_{i,\text{agent}}} \quad (12)$$

$$P_{\text{shared,agent}} = 1 - P_{\text{unique,agent}} \quad (13)$$

Deviations. To identify co-occurrences of schema pairs that are prevalent in an LLM’s language but not in human language, and vice versa, we calculate the level of deviation for each schema pair a, b as follow:

$$\Delta_{\text{Cooccurrence}_i} = |P_{\text{Cooccurrence}_i,\text{model}} - P_{\text{Cooccurrence}_i,\text{human}}| \quad (14)$$

where $P_{\text{Cooccurrence}_i,\text{model}}$ is the cooccurrence value of image schema pair i divided by the sum of cooccurrence values of all schema pairs for an LLM, $P_{\text{Cooccurrence}_i,\text{human}}$ is the cooccurrence value of image schema pair i divided by the sum of cooccurrence values of all schema pairs for humans. This allows us to identify the disproportionate co-occurrences encoded by a model with the greatest extent of deviation from human patterns.

5 Alignment Evaluation Results

Here we present an analysis of the sensorimotor regularities encoded by three contemporary LLMs: LLaMa-2-7b, LLaMa-2-70b, and gpt-4-1106. Our alignment evaluation examines their humanlikeness and identifies deviations from human patterns across three metrics including image schema distribution, conceptual associations, and image schema cooccurrence.

To contextualize the model–human alignment scores, we first established a human–human benchmark. We partitioned the human data into two demographically distinct subgroups—university-affiliated younger adults and older adults from a lifelong learning community—each consisting of 12 individuals. We then computed the human inter-group cosine similarity for all three metrics, and found substantial inter-group agreement for image schema distribution (0.9459), conceptual association (0.8605), and schema co-occurrences (0.6043). The high level of agreement between the two subgroups indicates that the sensorimotor regularities extracted from our human data are stable across diverse backgrounds. These benchmarks provide reference points for evaluating how closely each model’s alignment with humans approximates the level of agreement observed between human groups.

While large language models are trained on massive datasets potentially reflecting the language of millions of individuals, it is not feasible to construct a human baseline of comparable scale. Instead, our goal is to establish a robust and representative human baseline against which model outputs can be meaningfully compared. Because image schema–based patterns emerge from shared bodily experience and are theorized to be universal, a relatively small human sample can be sufficient for eliciting highly consensual patterns that reflect the conceptual processes of a broader population. The high level of agreement observed between our two demographically

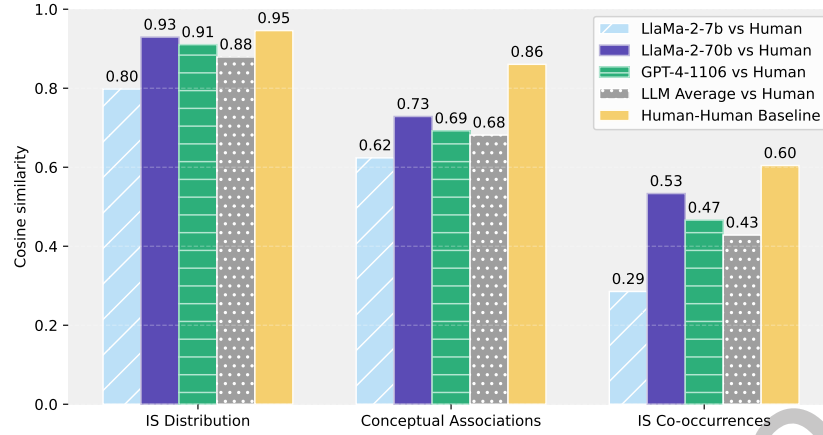


Fig. 3. Cosine similarity between sensorimotor patterns identified in model outputs and those identified in human responses. “LLM Average” represents the aggregate similarity score across all three LLMs relative to human patterns. “Human-Human Baseline” approximates the level of agreement observed between two demographically distinct human groups.

distinct subgroups confirms that the sensorimotor regularities extracted from our human data are stable across backgrounds, and therefore provide a suitable basis for meaningful comparison with patterns manifested in model outputs.

5.1 Image Schema Distributions

5.1.1 Humanlikeness.

Cosine Similarity. To compare the similarity between image schema distributions in a model’s and human language, we transform the two distributions into vector form (see Equation 1), and compute the cosine similarity between two vectors using Equation 2. As shown in Fig. 3, all three models closely mirror the image schema distribution in human language. Specifically, LLaMa-2-70b leads with a cosine similarity of 0.9294, followed by gpt-4-1106 at 0.9103, and LLaMa-2-7b at 0.7981. The human-human alignment value of image schema distribution is 0.9459. LLaMa-2-70b and gpt-4-1106 approximate this benchmark closely, indicating strong alignment in schema distribution patterns. In contrast, LLaMa-2-7b shows a more noticeable divergence, suggesting disproportionate representation of certain schemas relative to human patterns. The specific distributions of image schemas in models’ and human language are reported in Table 5 and visualized in Fig. 4.

Proportion of Unique and Shared Image Schemas. In addition to using cosine similarity as a global measure of similarity between human schema distribution and that of each model, we analyze the data from humans and from each model to determine the proportions of schemas that are unique to humans, unique to a particular model, and shared between humans and that model (see Fig. 5). Results indicate that all three models replicate a large range of human image schemas (88.03%–98.46%). Additionally, in all three models’ image schema distributions, only a small proportion of image schemas (1.54%–11.97%) are uniquely employed by the models and not present in human data (see Table 5).

Interestingly, although LLaMa-2-70b replicates the smallest range of human image schemas among all models, it reaches the highest cosine similarity with human data. This occurs because the model reproduces the most prevalent human schemas, but fails to capture many of the less common schemas that are nonetheless present

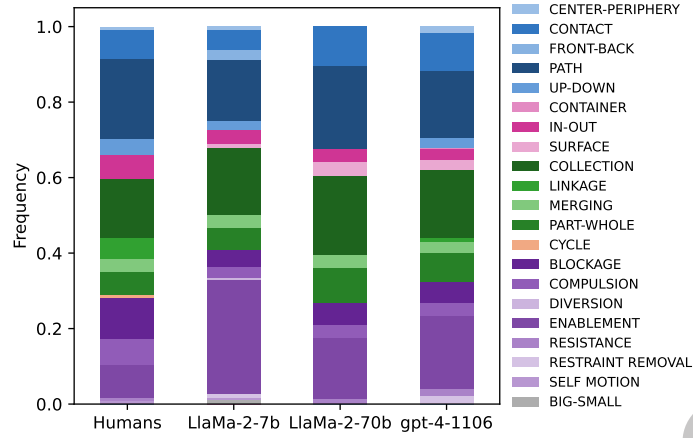


Fig. 4. Image schema distributions in human language and the linguistic outputs of three LLMs.

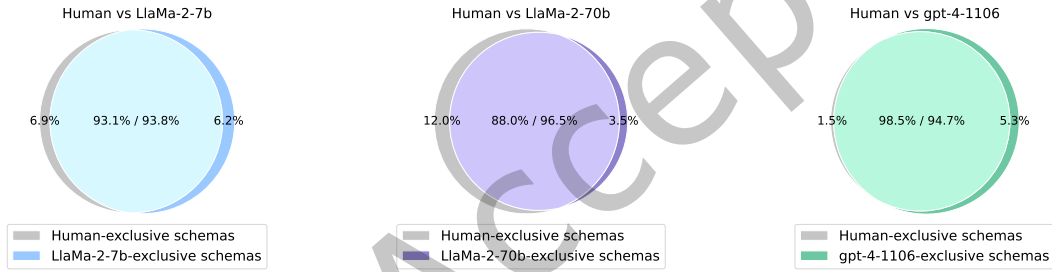


Fig. 5. The percentage of image schemas unique to humans, unique to a particular model, and shared between humans and that model.

in human data. While gpt-4-1106 replicates the largest proportion of human image schemas, it is surpassed by LLaMa-2-70b in terms of overall similarity. This discrepancy arises because it also uniquely produces a larger proportion of image schemas that are not found in human data.

5.1.2 Deviations. Despite the high similarity between image schema distributions in the models and in human language, we observe that each of the models tends to overuse or underuse certain image schemas compared to typical human patterns (see Fig. 6).

The three models exhibit similar patterns in their overuse or underuse of image schemas. For all three models, the image schema most excessively used in comparison to human data is **ENABLEMENT**, and all of them tend to also overuse **COLLECTION** and **SURFACE**. In terms of underuse, all models use **LINKAGE**, **BLOCKAGE**, **COMPULSION**, and **IN-OUT** less frequently compared to humans.

Apart from these similarities, each model demonstrates their unique patterns. For example, LLaMa-2-7b tends to overuse **FRONT-BACK** and **RESTRAINT REMOVAL**, while underusing **PATH**. LLaMa-2-70b tends to overuse **PART-WHOLE** and **CONTACT**, while underusing **UP-DOWN**. And gpt-4-1106 tends to overuse **CONTACT** and **RESTRAINT REMOVAL**, while underusing **PATH**.

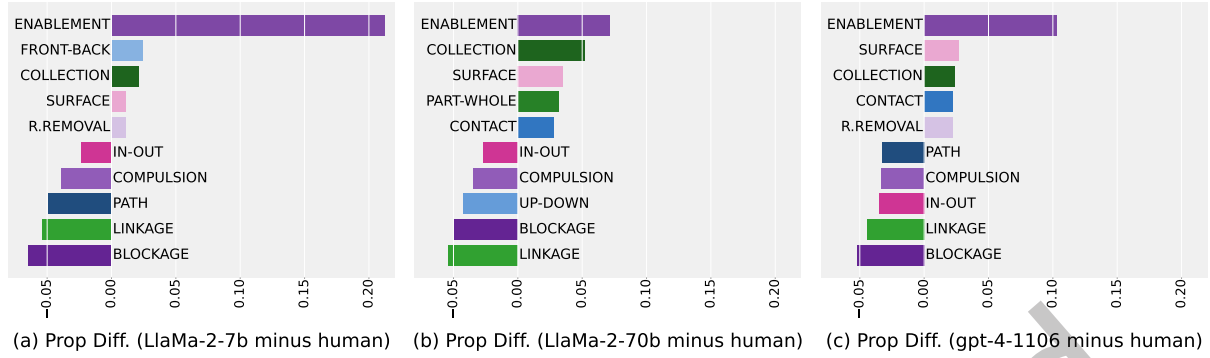


Fig. 6. Image schemas with notable proportional differences. A positive proportional difference indicates that the LLM overuses a schema compared to human data, while a negative proportional difference suggests that the LLM underuses a schema commonly found in human data. RESTRAINT REMOVAL is abbreviated to R.REMOVAL.

5.2 Conceptual Associations

5.2.1 Humanlikeness.

Cosine Similarity. To compare the similarity between conceptual associations between image schemas and concepts made by each model and humans, we transform the two distributions into vector form (see Equation 6), and compute the cosine similarity between two vectors.

As shown in Fig. 3, when it comes to conceptual associations, these models demonstrated decreased humanlikeness compared to mere image schema distributions. Again, LLaMa-2-70b leads with a cosine similarity of 0.7285, followed by gpt-4-1106 at 0.6922, and LLaMa-2-7b at 0.6238. The human-human alignment value for conceptual associations is 0.8605. Although LLaMa-2-70b outperforms other models, it still falls short of the benchmark, indicating a partial but incomplete replication of schema-concept mappings. LLaMa-2-7b and gpt-4-1106 similarly demonstrates moderate alignment, with notable divergences from human association patterns. The specific conceptual associations made by each model and human participants are summarized in Table 6 and visualized in Fig. 7.

Proportion of Unique and Shared Conceptual Associations. An analysis of each human-model reveals the proportions of conceptual associations that are unique to humans, unique to the model, and shared between both (see Fig. 8). While the three models effectively replicate a broad range of human image schemas, there is a noticeable drop in the range of human conceptual associations they capture, declining to between 76.06% to 83.01%. Moreover, all models exhibited a more pronounced tendency to employ uniquely invented conceptual associations. The proportion of unique conceptual associations among all conceptual associations created by each model varies from 19.83% to 33.42%.

Results indicate that gpt-4-1106 captures the broadest range of human conceptual associations (83.01%), slightly surpassing LLaMa-2-70b (77.22%). However, gpt-4-1106 is outperformed by LLaMa-2-70b in overall cosine similarity. This is because gpt-4-1106 employs a remarkably bigger proportion of unique conceptual associations that do not exist in human data, in comparison to LLaMa-2-70b. Some of these non-human-like conceptual associations employed by gpt-4-1106 occur frequently in the model's data, which substantially impacts the overall direction and magnitude of its conceptual association vector, resulting in a lower cosine similarity to human data.

5.2.2 Deviations. Fig. 9 shows the conceptual associations that the models use notably more frequently than humans, and those they use notably less frequently.

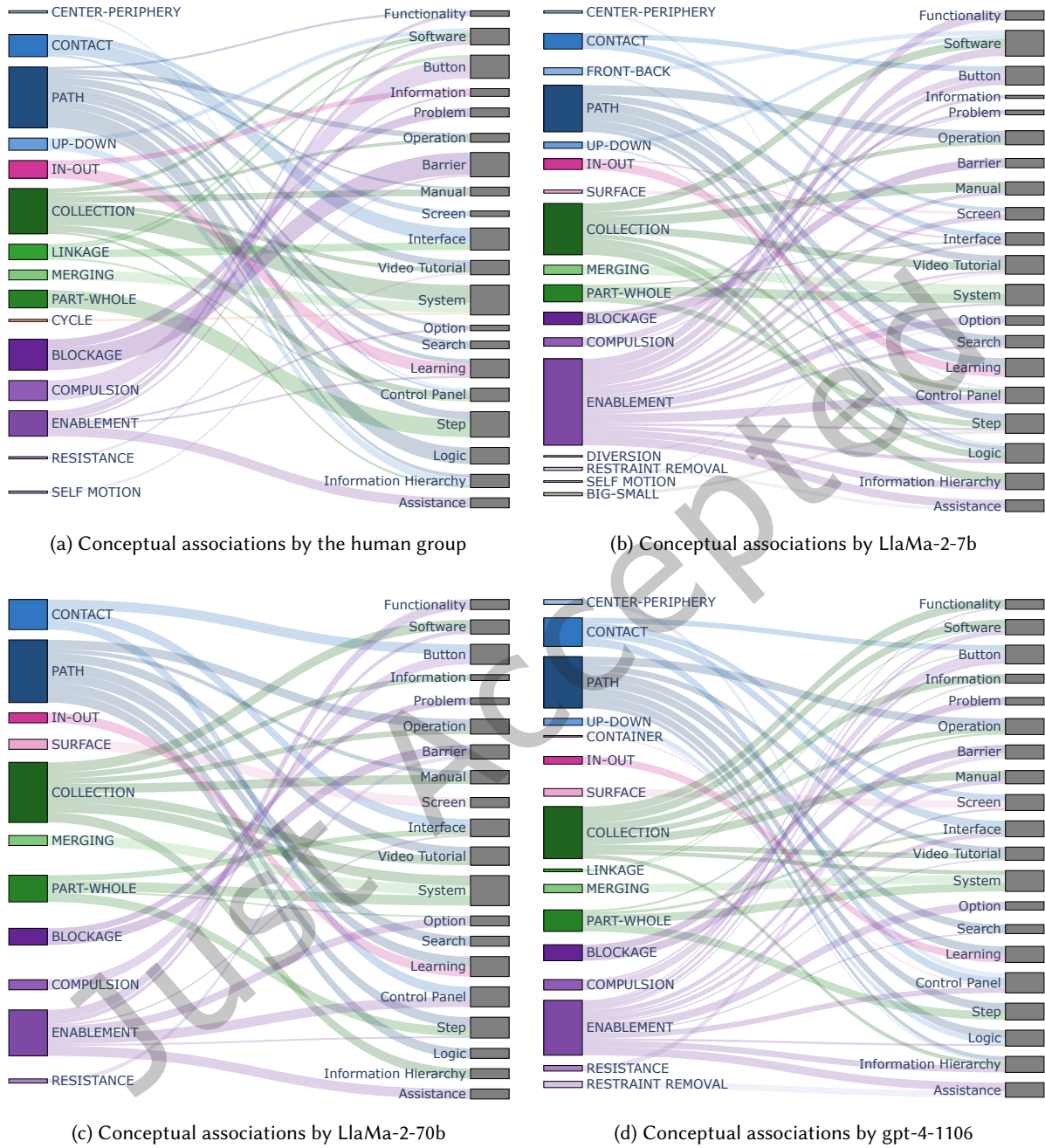


Fig. 7. Sankey diagrams showing conceptual associations by humans and three different models. Nodes on the left represent image schemas, and nodes on the right represent concepts.

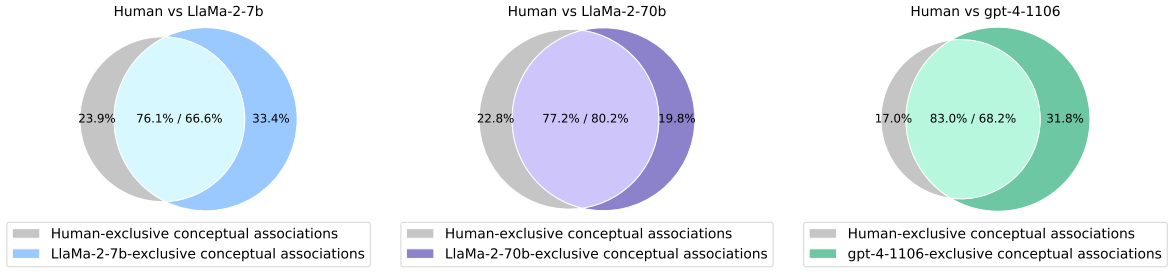


Fig. 8. The percentage of conceptual associations unique to humans, unique to a particular model, and shared between humans and that model.

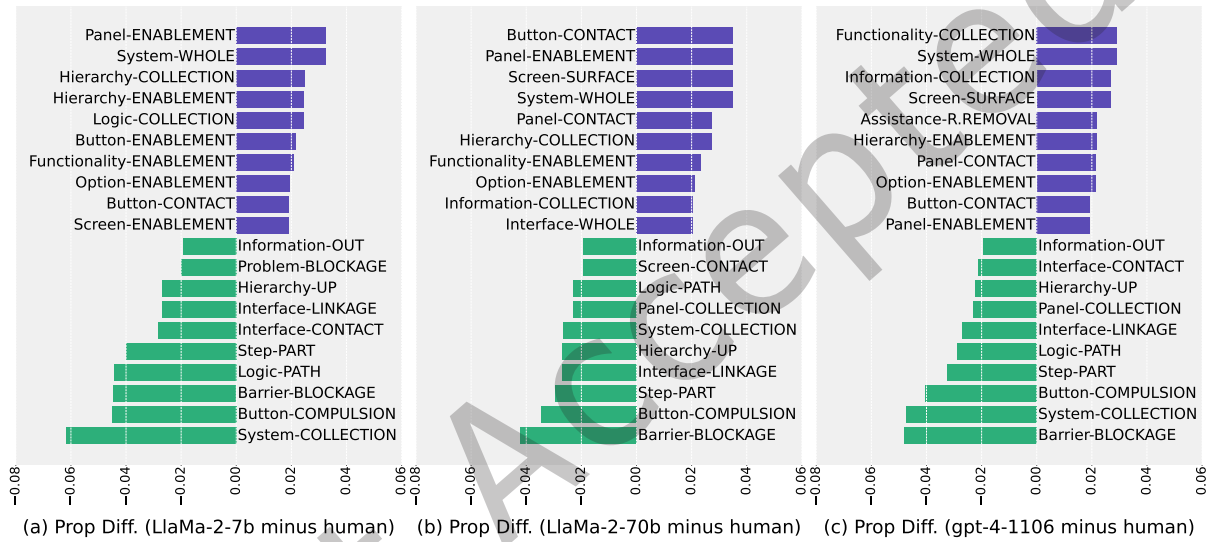


Fig. 9. Conceptual associations with notable proportional differences. A positive proportional difference indicates that the model employs a conceptual association more frequently than humans, while a negative proportional difference suggests that the model underuses an association prevalent in human data. For brevity in the plot, we simplified *Information Hierarchy* to *Hierarchy*, *Control Panel* to *Panel*, and *RESTRAINT REMOVAL* to *R.REMOVAL*.

The three models exhibit diversity in their overuse of certain conceptual associations, with a few consistent patterns. Notably, all models demonstrate a stronger tendency than humans to associate *System* with *WHOLE*, *Option* with *ENABLEMENT*, *Control Panel* with *ENABLEMENT*, and *Button* with *CONTACT*.

Interestingly, the three models demonstrate more consistent patterns in their underuse of certain conceptual associations that frequently appear in human data. All models show a weaker tendency than humans in associating *Information* with *OUT*, *Interface* with *LINKAGE*, *General Information* with *UP*, *Step* with *PART*, *Barrier* with *BLOCKAGE*, *Button* with *COMPULSION*, *Logic* with *PATH*, and *System* with *COLLECTION*.

5.3 Image Schema Co-occurrences

5.3.1 Humanlikeness.

Table 2. The co-occurrence value of the top 10 image schema pairs with the highest average co-occurrence discrepancies between human data and model data. Discrepancy is computed as the mean absolute difference between the human co-occurrence value and those of the three models (Llama-2-7b, Llama-2-70b, and gpt-4-1106).

Image Schema Pair	Human	Llama-2-7b	Llama-2-70b	gpt-4-1106
ENABLEMENT-COLLECTION	20	436	60	260
ENABLEMENT-CONTACT	0	123	168	247
COLLECTION-PATH	61	207	253	204
WHOLE-MERGING	0	144	144	144
COMPULSION-CONTACT	0	63	144	96
PATH-IN	44	144	144	132
CONTACT-LINKAGE	95	0	0	16
ENABLEMENT-PATH	20	130	24	130
COLLECTION-WHOLE	0	0	144	72
COLLECTION-MERGING	144	0	144	72

Cosine Similarity. Transforming the two distributions into vector form (see Equation 11) computing the cosine similarity between two vectors, allows a comparison of the similarity between image schema co-occurrences encoded by each model and in human language.

As shown in Fig. 3, the three models demonstrate varying degrees of performance decline when mirroring the human distribution of image schema co-occurrences, compared to their capabilities in replicating human-like image schema distribution and conceptual associations. LLaMa-2-70b leads with a cosine similarity of 0.5334, followed by gpt-4-1106 at 0.4661, and LLaMa-2-7b at 0.2855. The baseline human-human alignment for image schema co-occurrence is 0.6043. LLaMa-2-70b comes closest to this benchmark, though the remaining gap suggests an incomplete replication of combinatorial schema dynamics. gpt-4-1106 demonstrates some sensitivity but remains noticeably below the benchmark. LLaMa-2-7b falls substantially short, indicating limited capacity to model relational schema structures. Table 2 shows the top 10 image schema pairs with the greatest co-occurrence discrepancies between human data and model data. A complete summary of all image schema co-occurrences encoded by each model and in human language is provided in Table 7 and visualized in Fig. 10.

Proportion of Unique and Shared Schema Co-occurrences. An analysis of each human-model pair reveals the proportion of image schema co-occurrences distribution that are unique to humans, unique to the model, and shared between both (see Fig. 11). While all three models replicate a reasonable range of human conceptual associations, both LLaMa-2-7b and LLaMa-2-70b show performance declines in mirroring the human distribution of image schema co-occurrences. Specifically, LLaMa-2-7b's performance drops to 38.02%, and LLaMa-2-70b's to 52.56%. LLaMa-2-7b fails to capture seven out of ten schema co-occurrences most frequent in human data, and LLaMa-2-70b misses five out of ten most frequent human co-occurrences. In contrast, gpt-4-1106 demonstrates a robust performance, replicating 81.88% of the human distribution and capturing eight out of ten most frequent co-occurrences in human data. However, all three models demonstrate similarly high tendency to invent non-human-like co-occurrences (which can be considered a form of model confabulation/hallucination), ranging from 45.94% to 48.06% of their co-occurrences distribution.

Although gpt-4-1106 replicates the largest proportion of the human distribution of image schema co-occurrences, it is outperformed by LLaMa-2-70b in terms of overall cosine similarity to human distribution. This discrepancy arises because gpt-4-1106 excessively generates certain co-occurrences that, although present, are not frequent in the human distribution. Additionally, it generates 17 unique co-occurrences not found in human data, some

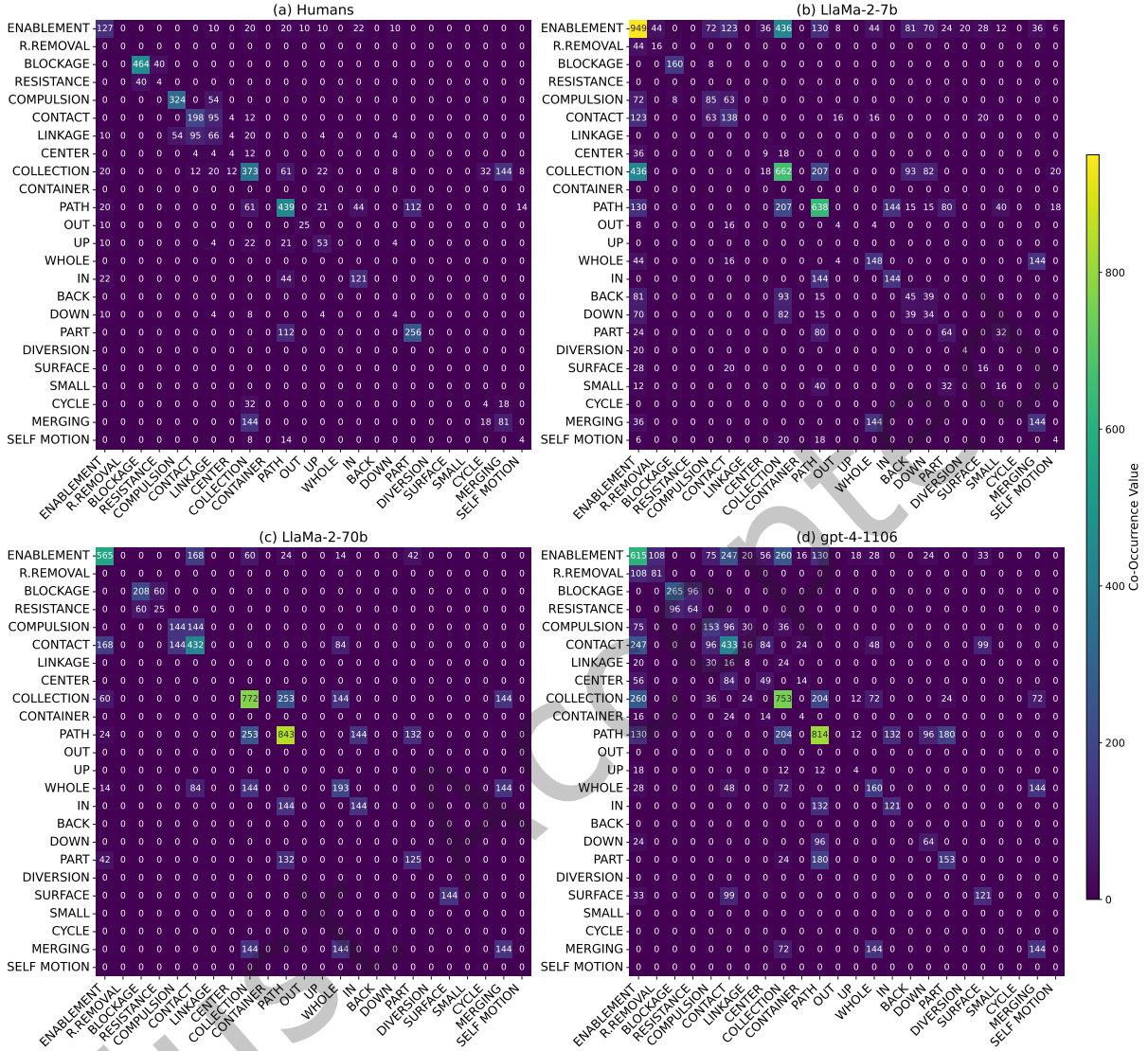


Fig. 10. Image schema co-occurrences found in (a) human language, (b) LLaMa-2-7b’s language, (c) LLaMa-2-70b’s language, and (d) gpt-4-1106’s language. The co-occurrence value between a pair of schema is calculated using Equation 10. A high co-occurrence value between two image schemas indicates that they frequently co-occur in the representation of the same concept. For brevity in the plot, we simplified RESTRAINT REMOVAL to R.REMOVAL.

with significant co-occurrence values (e.g., ENABLEMENT-CONTACT). These non-human-like, heavily weighted co-occurrences significantly affect the direction and magnitude of its co-occurrence vector, resulting in a lower cosine similarity to human distribution.

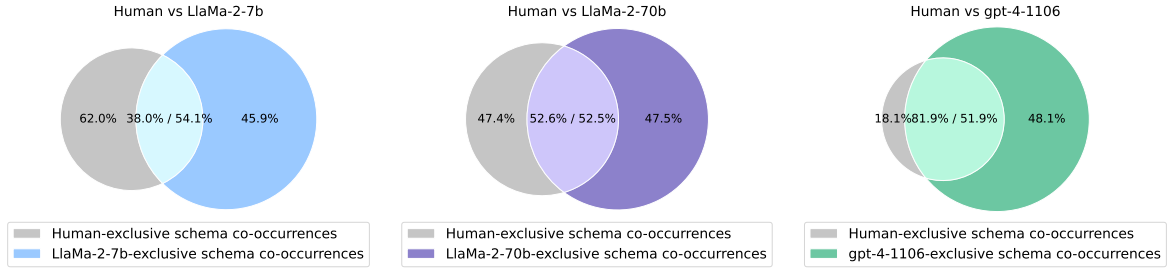


Fig. 11. The percentage of image schema co-occurrences unique to humans, unique to a particular model, and shared between humans and that model.

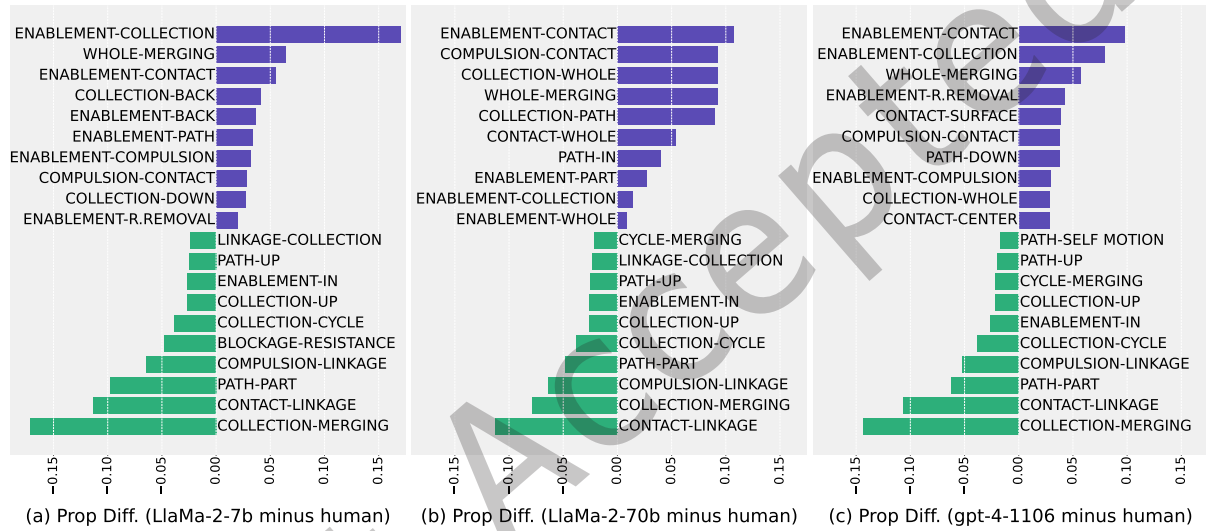


Fig. 12. Image schema co-occurrences with notable proportional differences. A positive proportional difference indicates that the LLM overuses a co-occurrence compared to human data, while a negative proportional difference suggests that the LLM underuses a co-occurrence commonly found in human data. For brevity in the plot, we simplified RESTRAINT REMOVAL to R.REMOVAL.

LLaMa-2-7b, on the other hand, can only replicate 38.02% of the human distribution. Additionally, it produces 25 unique co-occurrences not present in human data, with some being heavily weighted. This combination accounts for why LLaMa-2-7b attains the lowest overall cosine similarity (0.2855) to the human distribution of image schema co-occurrences.

5.3.2 Deviations. Fig. 12 shows image schema co-occurrences that each model uses notably more frequently than humans, and those each model uses notably less frequently.

Certain schemas co-occur under the same concept more frequently in the language of all three models than they do in human language. For example, all models demonstrate a stronger tendency than humans to concurrently use the schemas MERGING with WHOLE (i.e., two or more entities fitting together co-occurs with a unified whole with parts arranged in a particular fashion), ENABLEMENT with CONTACT (i.e., having the power to perform some act

co-occurs with the physical coming together of two or more things), *ENABLEMENT* with *COLLECTION* (i.e., having the power to perform some act co-occurs with a group of similar but autonomous objects), and *COMPULSION* with *CONTACT* (i.e., an entity forcing some passive entity to move co-occurs with the physical coming together of two or more things).

Interestingly, the three models show similar patterns in their failure to replicate the frequent co-occurrences of certain image schema pairs found in human data. All models show a weaker tendency than humans to concurrently use the schemas *PATH* with *UP* (i.e., following a path co-occurs with moving upward), *ENABLEMENT* with *IN* (i.e., having the power to perform some act co-occurs with something entering or being inside an area), *COLLECTION* with *UP* (i.e., a group of similar but autonomous objects co-occurs with moving upward), *COLLECTION* with *CYCLE* (i.e., a group of similar but autonomous objects co-occurs with a cycle that loops through events and restarts), *COMPULSION* with *LINKAGE* (i.e., an entity forcing some passive entity to move co-occurs with a link connecting two or more entities), *PATH* with *PART* (i.e., following a path co-occurs with a single part of a larger whole), *CONTACT* with *LINKAGE* (i.e., the physical coming together of two or more things co-occurs with a link connecting two or more entities), and *COLLECTION* with *MERGING* (i.e., a group of similar but autonomous objects co-occurs with two or more entities fitting together to form a whole).

6 Downstream Evaluation

While the previous section evaluated human-LLM alignment at conceptual level via sensorimotor regularities, it remains an open question whether such alignment translates into practical benefits in downstream use cases. To investigate this, we conducted a follow-up study examining the downstream effects of incorporating human-aligned sensorimotor knowledge into gpt-4-1106 via prompt-based intervention. This targeted intervention involved injecting image schema co-occurrence patterns that were frequently observed in human conceptualizations but underrepresented in the model’s original outputs. This allows us to assess whether bringing the model into closer alignment with human sensorimotor regularities improves model performance from an end-user perspective.

6.1 Context

Evaluating the downstream effects of additional sensorimotor knowledge in LLMs is methodologically nontrivial, as many standard tasks can fail to reveal its distinctive contributions. Pre-trained LLMs possess extensive prior linguistic representations, and tasks such as generating instructions for a product typically rely on well-represented patterns already internalized during pre-training, making it difficult to isolate the effect of additional sensorimotor knowledge.

To meaningfully assess its impact, we turned to creative writing, a domain that allows for the introduction of novel, fictional concepts (e.g., *mayaki*, *kengchuu*) that lie outside the model’s pretraining distribution and even beyond existing human knowledge. We presented the model with opening clauses that introduce fictional terms while evoking an image schema—for example, “*The mayaki is planted inside the kengchuu,*” (evoking the *IN* schema)—and asked the model to generate a continuation.

6.1.1 Hypotheses.

Hypothesis 1: We hypothesize that, when augmented with human-aligned sensorimotor knowledge, the model will generate a continuation that evoke a second image schema that frequently co-occur with the schema evoked by the prompt clause, following frequently observed human patterns (H1). For example, generating a continuation evoking *ENABLEMENT* when given a prompt clause evoking *IN*. Without the additional sensorimotor patterns, the baseline model—when presented with prompt clauses evoking image schemas—is neither expected to systematically generate continuations based on such schema co-occurrence patterns nor rely on shortcuts from

training data, since the fictional terms lie outside its pretraining distribution. This limits its ability to reproduce human-like schema associations purely from memorized patterns.

Hypothesis 2: The second hypothesis is that the sentence continuations generated by the model augmented with human-aligned sensorimotor knowledge will be perceived by human evaluators as conceptually clearer, more contextually contingent, and more human-like compared to continuations generated by the baseline model (H2). Drawing on the work of Johnson [41], image schemas “can be recruited to structure abstract concepts and to carry out inferences about abstract domains of thoughts”. Many abstract concepts become clear and intelligible through their grounding in bodily schemas, for example, understanding time as a *PATH*. Therefore, we hypothesize that sentence continuations structured around image schemas will provide human evaluators a clearer understanding of the abstract concepts described. Moreover, when the relationship between the open clause and the continuation is structured around a pair of image schemas that frequently co-occur in human conceptualizations, this schema pair provides a familiar conceptual structure that meaningfully ties the continuation to the open clause. As a result, such continuations should be perceived as more contextually contingent, that is, as more clearly dependent on, or resulting from, the open clause. Finally, image schema-based scaffolding is a fundamental feature of human embodied reasoning. When instructed to follow these scaffolds, a model’s outputs should more closely approximate the structure and logic of human language production. This motivates our prediction that schema-guided continuations will also be perceived as more likely written by a human.

Hypothesis 3: We further hypothesize that sensorimotor-guided continuations will not be judged as less imaginative than those generated under the baseline prompt by humans (H3). This hypothesis draws on prior research suggesting that image schema-based scaffolding can simultaneously support intuitive understanding and perceived novelty [38, 48].

6.2 Prompt-Based Sensorimotor Alignment

We implemented two prompt variants for gpt-4-1106: one embedding additional human-aligned sensorimotor patterns, and one entirely schema-free.

6.2.1 Baseline Prompt. The baseline prompt (see Listing 1) contained only general writing instructions, without any schema content.

You are a creative writing assistant. Your task is to respond with a continuation when given a partial sentence.
 Maintain and respect the given terms---do not ‘correct’ or reinterpret made-up words.
 Use creative and atmospheric language that evokes a vivid sense of world and mood.
 Avoid long or elaborate continuations. Keep the completion concise---roughly similar in length to the opening clause.
 Do not produce any new sentences beyond the completed sentence.

6.2.2 Sensorimotor-Guided Prompt. In the sensorimotor-aligned prompt (see Listing 2 for the complete prompt), we preserved the same writing instructions as in the baseline but augmented them with human-aligned image schema co-occurrence patterns to steer gpt-4-1106’s behavior. Beyond the shared writing instructions, the prompt comprises three key components:

- **Schema-guided instructions.** Grounded in empirical findings from our analysis of image schema co-occurrences patterns (Section 5.3.2), the prompt includes a ranked list of image schema pairs, each pair assigned a numerical weight quantifying their underrepresentation in gpt-4-1106’s outputs relative to human data. Each sentence completion is required to be conceptually structured around one of these schema pairs, meaning that the completed sentence should instantiate the meanings of both schemas and

express their relationship. Higher-weighted schema pairs are preferentially considered, but the prompt allows flexible adaptation: if the top pair does not fit the input context, the model may select the next most appropriate pair. The ranked list of schema pairs are shown below:

- COLLECTION and MERGING: 0.143
- CONTACT and LINKAGE: 0.107
- PATH and PART: 0.062
- COMPULSION-LINKAGE: 0.053
- CYCLE-COLLECTION: 0.038
- ENABLEMENT-IN: 0.026
- COLLECTION-UP: 0.021
- CYCLE-MERGING: 0.021
- PATH-UP: 0.020
- PATH and SELF MOTION: 0.017

- **Image schema knowledge base.** To support the interpretation and selection of schema pairs, the prompt includes a compact dictionary defining each image schema in the ranked list. For example:

3. CONTACT: Denotes physical coming together of two or more things.
 4. LINKAGE: Denotes two or more entities which are connected with each other by means of a linking device of some kind.
 CONTACT and LINKAGE are usually used to describe the same concept.

11. ENABLEMENT: Denotes having the physical or metaphorical power to perform some act.
 12. IN: Denotes the location of an entity within a CONTAINER or bounded area.
 ENABLEMENT and IN are usually used to describe the same concept.

- **Few-shot examples.** The prompt includes representative input–output pairs for each schema pair. The input shows the opening clause to be completed; the output demonstrates how that clause can be meaningfully extended using the guiding schema pair as a conceptual scaffold. For instance:

CONTACT and LINKAGE
 Input: The two wires are attached at the ends,
 Output: The two wires are attached at the ends, allowing electricity to pass through their connection.

ENABLEMENT and IN
 Input: The nutrients are absorbed into the intestine,
 Output: The nutrients are absorbed into the intestine, giving the body the energy it needs to function.

This lightweight but principled intervention enables a direct comparison between completions guided by sensorimotor regularities and baseline completions.

6.3 Sentence Continuations

Having established our prompt designs, we fed gpt-4-1106 an identical set of ten opening clauses—each designed to evoke a particular image schema—under both the sensorimotor-guided and baseline prompt conditions, and generated sentence continuations with a temperature of 0. This design employed a deterministic first-shot setup rather than multi-shot sampling, as temperature = 0 eliminates stochastic variance and multiple shots would have produced identical continuations. Deterministic generation is widely used in studies that evaluate the

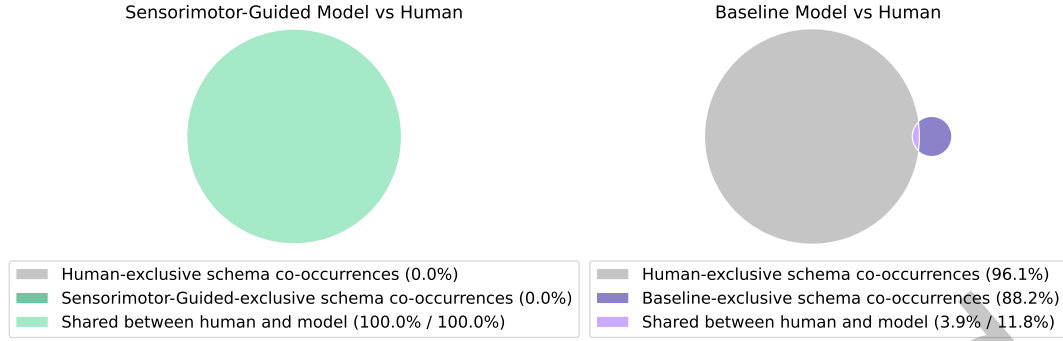


Fig. 13. Comparison of schema co-occurrence patterns between human data and model outputs before and after sensorimotor augmentation. The SRs-guided model (after augmentation, left) produced only schema co-occurrences also observed in human data, whereas the baseline model (before augmentation, right) generated 88.2% of co-occurrences not present in the human distribution. The SRs-guided model exhibited substantially greater overlap with human schema co-occurrence patterns than the baseline model.

downstream behavior and performance of LLMs (e.g., [42, 58, 65, 70]), as it ensures that any observed differences are attributable to the intervention rather than sampling variability.

Table 3 shows the resulting sensorimotor-guided and baseline sentence continuations side by side, with image schemas identified in each continuation. We then compared the SRs-guided and baseline outputs across three metrics (image-schema distribution, see Table 10 in Appendix; conceptual association, see Table 11 in Appendix; and schema co-occurrence, see Table 12 in Appendix) to examine how the augmentation shifts the sensorimotor patterns of generated outputs. The cosine similarity between SRs-guided and baseline outputs was 0.18 for image-schema distribution, 0.07 for conceptual association, and 0.13 for schema co-occurrence. Such low similarities indicates that the sensorimotor augmentation led to a substantial shift in the outputs' image-schematic patterns.

To assess the direction of this shift relative to human sensorimotor patterns, we compared both SRs-guided outputs and baseline outputs to human data on the schema co-occurrence metric, which captures relational regularities between image schemas independent of specific concepts. The SRs-guided model demonstrated substantially greater schema co-occurrence similarity to the human pattern (cosine = 0.71) than the baseline model, which demonstrated a strong deviation from human distribution (cosine = 0.04), indicating that the augmentation produces outputs whose schema co-occurrence patterns more closely reflects human sensorimotor regularities. As illustrated in Figure 13, the SRs-augmented model activated only schema co-occurrences that were also observed in human data, while 88.2% of the co-occurrences produced by the baseline model were absent from the human distribution.

This pattern supports Hypothesis 1, which predicted that augmenting the model with human-aligned sensorimotor knowledge would lead it to generate continuations that evoke second image schemas consistent with frequently observed human co-occurrence patterns. These results indicate that embedding human-aligned sensorimotor regularities into prompts can reliably steer the model toward conceptually grounded continuations that reflect human-like sensorimotor patterns, while unguided models tended to produce co-occurrences that deviated from such patterns.

6.3.1 Limitations. While this prompt design demonstrates that language models *can* be guided to produce sentence continuations grounded in a targeted set of ten image schema pairings, it does so through a deliberately controlled setup. The constraints serve the purpose of proof-of-concept and more importantly, a direct comparison

Table 3. Sentence continuations generated by gpt-4-1106 following prompts guided by sensorimotor regularities versus baseline prompts. As input, each open clause features fictional concepts (e.g., *mayaki* and *kengchuu*) and evokes an image schema (e.g., “*The mayaki is planted inside the kengchuu*,” evokes IN).

Open Clause	SRs-Guided Continuation	Baseline Continuation
<i>The Uokata emerged from the kikis’ reunion,</i> (Schema evoked: COLLECTION)	<i>blending their individual energies into a harmonious collective force.</i> (Schema evoked: MERGING)	<i>its scales shimmering with the iridescent tales of a thousand whispered secrets.</i> (Schema evoked: BRIGHT, COLLECTION)
<i>The mayaki is planted inside the kengchuu,</i> (Schema evoked: IN)	<i>enabling the kengchuu to flourish with vibrant hues.</i> (Schema evoked: ENABLEMENT)	<i>casting a shadow that dances with the whispers of the wind.</i> (Schema evoked: DARK)
<i>A meiki evolves through Dream, Mist, and then back to Dream,</i> (Schema evoked: CYCLE)	<i>completing a cycle that merges the ethereal with the tangible.</i> (Schema evoked: CYCLE, MERGING)	<i>each phase shimmering with the ethereal whispers of ancient lore.</i> (Schema evoked: BRIGHT)
<i>The panomi forced the rikis through the veins of the Kemathi,</i> (Schema evoked: COMPULSION)	<i>creating a linkage that spread the healing energy throughout its body.</i> (Schema evoked: LINKAGE)	<i>their whispers echoing like wind through ancient caverns, shaping destinies with unseen hands.</i> (Schema evoked: CONTAINER)
<i>The rikis leaned into the soft frame of the sleeping kengchuu,</i> (Schema evoked: CONTACT)	<i>establishing a gentle contact that linked the two creatures in peaceful repose.</i> (Schema evoked: CONTACT, LINKAGE)	<i>its breaths a gentle rhythm beneath the star-swathed sky.</i> (Schema evoked: DOWN)
<i>The Haya begins with Jenori and ends with Kanawa,</i> (Schema evoked: PATH)	<i>each village and town en route marking a segment of its winding journey.</i> (Schema evoked: PART, PATH)	<i>but in the twilight hours, it whispers secrets only the wind can truly understand.</i> (Schema evoked: DARK)
<i>The Thousand Reinos drifted up through the cracks in gravity,</i> (Schema evoked: UP)	<i>forming a rising collection of celestial bodies in the vastness of space.</i> (Schema evoked: COLLECTION)	<i>their silvery sails unfurling like whispers against the fabric of space-time.</i> (Schema evoked: PATH, CENTER-PERIPHERY)
<i>The teiki cycle begins with a renerity and ends with a zeltr,</i> (Schema evoked: CYCLE)	<i>completing a series of phases that merge into a continuous loop.</i> (Schema evoked: COLLECTION, MERGING, CYCLE)	<i>but in between, the world is awash with the luminescent blooms of the noctilucous jaspets.</i> (Schema evoked: BRIGHT)
<i>The reiki surged from the lowlands to the Himako,</i> (Schema evoked: PATH)	<i>its energy ascending along the sacred path to the temple’s peak.</i> (Schema evoked: UP, PATH)	<i>casting a shimmering veil over the ancient, whispering pines.</i> (Schema evoked: UP, BRIGHT)
<i>Without any command, the jupyteri moved on its own,</i> (Schema evoked: SELF-MOTION)	<i>tracing a spontaneous path across the digital landscape.</i> (Schema evoked: PATH)	<i>its tendrils undulating with a silent grace through the twilight mist.</i> (Schema evoked: SMOOTH, DARK)

between outputs generated with or without targeted conceptual scaffolding. However, practical implementations should adopt a more comprehensive knowledge base encompassing a wider range of schema pairings, and include fallback strategies for inputs that do not align well with any known schema pair. Such refinements would enhance both the robustness and ecological validity of sensorimotor-guided generation.

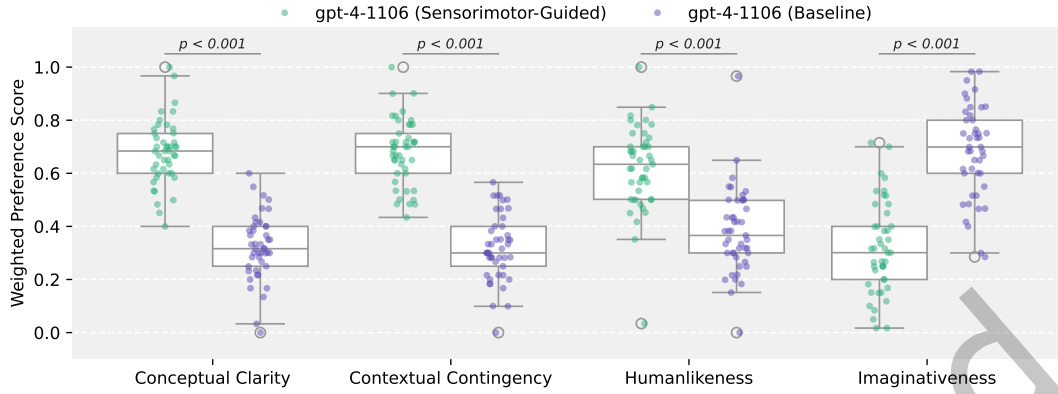


Fig. 14. Distributions of weighted preference ratings for sensorimotor-guided sentence continuations and baseline continuations along four dimensions. The box plots show the median (the horizontal line), the first and third quartile (the box), and the minimum and maximum (the whiskers). The circle signs (‘o’) indicate outliers. The scatter dots represent the mean weighted preference rating per participant for each condition and dimension.

6.4 Human Evaluation

To investigate Hypothesis 2 and 3 we carried out a within-subjects study with 50 fluent English speakers recruited on Prolific. In each of the ten trials, participants were shown an open clause alongside two continuations (See Table 3)—one sensorimotor-guided and one baseline—labeled as either *A* or *B*. The assignment of label *A* to a condition was counterbalanced. Participants then provided a weighted preference rating on a 7-point Likert scale (1 = Definitely *A*, 7 = Definitely *B*, 4 = Uncertain) along four dimensions: *Conceptual Clarity* (e.g., *Which completion gives you a better understanding of the concepts of “mayaki” and “kengchuu”, and how certain are you?*), *Contextual Contingency* (e.g., *Which completion feels more contingent on (i.e., resulting from) the opening clause, and how certain are you?*), *Imaginativeness* (e.g., *Which completion feels more imaginative, and how certain are you?*), and *Humanlikeness* (e.g., *Which completion seems more likely written by a human writer, and how certain are you?*).

We excluded participants with unusually long or short completion times using the standard interquartile range (IQR) method, removing those whose durations were either more than 1.5 times the IQR above the third quartile or below the first quartile. We also excluded two participants who failed the attention check. This procedure excluded 5 out of 50 participants, resulting in a final sample of 45 participants for analysis. The 45 participants ranged in age from 25 to 70 years ($M = 45.3$, $SD = 12.0$), with a balanced gender distribution (22 women, 22 men, and 1 non-binary participant). We analyzed participants’ ratings across four dimensions.

Weighted Preference Analysis. Each 7-point rating (1 = Definitely *A*, 7 = Definitely *B*, 4 = Uncertain) simultaneously captured both the direction and the confidence of a participant’s preference. We describe these as *weighted preferences* because the distance from the midpoint reflects how confidently a participant favored one continuation over the other. This approach differs from a binary (unweighted) preference measure, which records only the direction of choice without accounting for confidence weighting. Likert-scale responses were transformed into normalized scores between 0 and 1 and grouped by condition to reflect the degree of preference for the sensorimotor-guided continuations versus Baseline continuations. Normality assumptions (Shapiro-Wilk test) were met for all dimensions except *Humanlikeness*, for which we applied a non-parametric Wilcoxon signed-rank test.

Table 4. Statistical comparison of sensorimotor-guided and baseline sentence continuations across four dimensions.

Dimension	Test	$M_{\text{SRsGuided}}$	M_{Baseline}	Statistic	p	Effect Size
Conceptual Clarity	t -test	0.679	0.321	$t(44) = 9.90$	$< .001$	Cohen's $d = 1.477$
Contextual Contingency	t -test	0.680	0.320	$t(44) = 9.70$	$< .001$	Cohen's $d = 1.447$
Humanlikeness	Wilcoxon	0.620	0.380	$W = 92.5$	$< .001$	rank-biserial $r = 0.622$
Imaginativeness	t -test	0.318	0.682	$t(44) = -7.09$	$< .001$	Cohen's $d = -1.057$

As shown in Figure 14 and Table 4, participants rated the sensorimotor-guided sentence continuations significantly higher than baseline continuations for *Conceptual Clarity*, *Contextual Contingency*, and *Humanlikeness*, consistent with Hypothesis 2. Contrary to Hypothesis 3, participants rated the baseline continuations as significantly more imaginative.

Taken together, these results support Hypotheses 1 and 2 while disconfirming Hypothesis 3. The model augmented with targeted sensorimotor priors produced continuations that embedded schema co-occurrence patterns more closely aligned with human data (H1), and such closer alignment yielded downstream benefits in interpretability, coherence, and humanlikeness (H2). While prior research in interaction design suggests that image-schema-based scaffolding can simultaneously support intuitiveness and novelty, our downstream evaluation results point to a potential trade-off in LLM sentence continuation (H3). One possible explanation is that the sensorimotor-guided prompts introduced more explicit structure and constraints, which could reduce the flexibility of the model's expressions. Alternatively, the lower imaginativeness ratings may reflect that the baseline model produced more unconventional schema co-occurrence, as 88.2% of its schema co-occurrence patterns were not observed in human data, while the augmentation re-introduced conventional human-like schema pairings. In this sense, the sensorimotor augmentation traded imaginative divergence for conceptual conventionality, steering generation toward outputs that are more typical of human cognition. These results highlight that while sensorimotor-guided prompting could enhance interpretability, coherence, and humanlikeness, it may also impact the perceived creativity of outputs in subtle ways.

7 Discussion

7.1 A Human-like Surface with a Less-human Core

The results indicate that these language models are equipped with a set of highly human-like conceptual building blocks, as shown by their impressive capabilities in mirroring the distribution of image schemas found in human data. However, the models make use of these conceptual building blocks in a less human-like manner, as evidenced by their decreased similarities to the human distribution of conceptual associations. Further, the models exhibit patterns of schema entailment that significantly differ from those observed in human data, as indicated by even lower similarities to the human distribution of image schema co-occurrences. As discussed in section 2.2, these three metrics represent progressively sophisticated processing of image schemas. Beneath the surface of producing outputs that reflect highly human-like sensorimotor gestalts and their corresponding functions, these models exhibit limited capability in combining these gestalts to support human-like conceptual reasoning.

On the macro level, one possible reason for the sensorimotor gap observed between human and model concept representations lies in the nature of the models' training data. Our human benchmarks, elicited through one-on-one interviews, capture participants' spontaneous, intuitive efforts to contextualize and make sense of abstract concepts—responses that likely represent the cognitive strategies humans intuitively activate during real-time conceptualization. In contrast, the training data of large language models consists predominantly of written text, which limits their exposure to the embodied, ad-hoc reasoning typically found in spontaneous speech. This modality gap may partially explain why, when given the same set of interview questions, LLMs failed to

replicate the sensorimotor patterns observed in human responses. These intuitive structures, embedded in human real-time conceptualization but currently inaccessible to LLMs, are particularly valuable for improving human-LLM interaction, as they offer a foundation for building models whose outputs can align more naturally with human meaning-making processes in novel or evolving situations. This interpretation aligns with recent work by Cuskley et al. [16], who argue that the current pre-training that exposes large language models exclusively to unimodal textual data (primarily written language), can lead to models that encode only surface-level functional patterns of human *writing*, without being about to represent the contextual, multimodal and interactional aspects of human linguistic use and meaning-making. This reinforces the view that such modality limitations constrain LLMs' ability to approximate the mechanisms underlying human linguistic behavior and cognition. Recognizing this modality gap highlights promising directions for curating training data and developing targeted fine-tuning strategies to enhance alignment between model outputs and human cognition.

7.2 Similarities and Differences in Non-human Traits

We observe certain commonalities in all three models' deviations from typical human patterns, which may stem from the common limitations in their training data or training objectives. For example, all three models frequently overuse the image schema of *ENABLEMENT* and *COLLECTION*. The former implies a bias toward casual framing, representing concepts as initiating for enabling other phenomena or actions. The latter reflects a grouping bias, representing concepts as groupings of similar objects. In terms of conceptual associations, some frequent associations in the human distribution are less frequent in the distributions of all three models. One example is that humans often metaphorically conceptualize *Logic* as a *PATH* (e.g., "*Logic is moving from one step to another in sequence.*", "*It (logic) is a well-defined sequence of events, which lead from a point A to point B.*", "*It (logic) is a route where you arrive at a conclusion from a set of starting points*"), whereas all models generally show a weaker tendency to make this association (e.g., Llama-2-7b: "*(Logic is) methods used to reason and draw conclusions.*"; Llama-2-70b: "*It (logic) can be thought of as the 'thinking' behind a technology.*"; gpt-4-1106: "*Logic is the systematic method of reasoning.*"). This discrepancy could be attributed to language models' reliance on statistical regularities in text, coupled with the occasional weak representation of metaphorical connections between concepts and image schemas in the linguistic training data. When it comes to mirroring human distribution of image schema co-occurrences, the three models demonstrate similarities in their weaknesses. For example, all models demonstrate a weaker tendency than humans to concurrently use *ENABLEMENT* with *IN* to represent the same concept. The *ENABLEMENT-IN* co-occurrence was instead frequently observed in human data. The concept of *Learning* is jointly structured by both schemas across different expressions. For example, *ENABLEMENT* is evoked in statements such as "*(Learning is) to enable a gap in your understanding to be filled.*" and "*(Learning is) to be able to repeat the situation once you know how to do something*", which describe learning as a process that empowers further actions or understanding. Meanwhile, *IN* is evoked in expressions such as "*Learning is taking in information or experiences*" and "*learning is when you receive more information and you internalize it*", which frames learning as internalization of knowledge. Taken together, these linguistic examples illustrate that learning is conceptualized as a process that both internalize new information (*IN*) and enables subsequent subsequent actions (*ENABLEMENT*). This human bias can be attributed to humans having the sensorimotor experiences of internal components—such as a heart in a body or a processor in a computer—enabling the body or the computer to work, while the language models do not.

Through identifying the percentage distributions of unique and shared patterns between humans and a model, we examined the behavioral traits that contribute to a model's success or failure in mimicking human distributions. For example, Llama-2-70b represents a more conservative and steady profile, achieving the high overall human-likeness by replicating a modest range of human distributions while generating the very few non-human-like patterns. A model with these traits might be less likely to surprise human users during initial interactions, easier

to predict and manage. However, as interactions progress, human users might encounter difficulties in aligning objectives with such a model, particularly when these objectives require human sensorimotor knowledge that the model has not replicated. In contrast, gpt-4-1106 represents a capable yet more chaotic profile, replicating the broadest range of human patterns but also generating a significant proportion of unique patterns that are absent in human data. A model with such traits might sometimes appear more unpredictable and difficult to manage for human users. However, the good coverage of human sensorimotor regularities gives the model substantial potential to align with human objectives across a wide range of scenarios, provided that the prompts are thoughtfully crafted.

7.3 Towards Human-like Sensorimotor Encoding

The models evaluated in this study are of varying sizes. A natural question is whether the humanlikeness of sensorimotor regularities encoding improves as model parameters increase. We observed enhancements in humanlikeness from a smaller model to a larger one, with LLaMa-2-70b outperforming LLaMa-2-7b across all three metrics. However, beyond a certain point, additional parameters may not yield proportional improvements, as shown by gpt-4-1106, which has more parameters yet did not outperform LLaMa-2-70b in any of the three metrics in terms of overall similarity. Therefore we conjecture that merely more parameters might not guarantee greater overall humanlikeness in the encoding of sensorimotor regularities.

To enhance humanlikeness of sensorimotor regularities encoding in language models, targeted fine-tuning or in-context learning strategies might be necessary. Here we take the three models in our alignment evaluation as an example, to discuss how the evaluation results can inform the design of training data and training objectives of the targeted fine-tuning. All three model are equipped with the “correct” (or highly human-like) set of image schemas, and are able to associate at least one correct image schema to each concept. Such models should be further trained to predict other schemas that might logically co-occur with a this concept based on human intuition about schema co-occurrences. To achieve this, the training data should focus on promoting human-like image schema co-occurrences while actively filtering out the non-human-like co-occurrences uniquely invented by the model. The training objective should focus on enhancing the model’s ability to predict other related image schemas for a given concept, based on one or more known image schemas associated with that concept. The effectiveness of such fine-tuning can be assessed through the re-trained model’s ability to establish highly human-like conceptual associations for concepts not present in the training set. A promising direction for future research involves targeted fine-tuning to enhance humanlikeness of sensorimotor regularities encoding in language models, based on the deviations from human patterns highlighted by this framework.

Beyond model-level tuning, our downstream evaluation demonstrates that targeted human-aligned sensorimotor priors can also be leveraged at the prompt level to improve output quality. By guiding gpt-4-1106 with targeted schema co-occurrence patterns derived from its observed misalignment with human patterns, we elicited sentence continuations that were rated by human evaluators as conceptually clearer, more contextually contingent, and more human-like, compared to those generated by the same model without access to these targeted priors. This result highlights that even lightweight, prompt-based interventions can yield measurable benefits in downstream use cases when grounded in insights gained from sensorimotor alignment evaluation. Although such prompt-based interventions may not offer the generalizability that could be achieved through fine-tuning, they provide an efficient and scalable mechanism for improving the sensorimotor alignment and quality of model outputs in structured settings where the conceptual constraints of a tasks are well defined. Future studies should examine the downstream impacts of targeted sensorimotor-guided prompting in tasks beyond creative writing, to further assess its potential for improving sensorimotor alignment and output quality across diverse task domains. A promising direction for future research is to extend the downstream evaluation across multiple base models with differing initial levels of sensorimotor alignment with humans. Applying the

same augmentation procedure to several models and comparing their pre- and post-augmentation performance would allow for further examination of how a model’s initial sensorimotor alignment influences its downstream performance as well as how it responds to targeted sensorimotor augmentation.

As discussed in section 2.2, when a language model successfully encodes human intuition about image schema co-occurrences, it can more accurately infer multimodal properties of a concept (e.g., spatial relations) from textual information alone. Another intriguing area for future research would be to investigate whether a language model fine-tuned to better align with human sensorimotor regularities encoding, exhibit improved performance in tasks that require cross-modal reasoning (e.g., spatial understanding).

7.4 Limitations

We list several limitations in our alignment evaluation study. First, to establish a human baseline for our empirical demonstration of the framework, we conducted interviews with a fixed sample of 24 participants (half aged 21–34, recruited from a large university in the UK; half aged 62–75, recruited from a UK educational charity for people who are not or no longer in full-time employment). While the strong agreement observed between these two age groups suggests that our human benchmark is robust and stable across demographically different participants, future iterations of the study would benefit from broader demographic diversity, such as including participants with varying levels of educational attainment.

Moreover, we collected the human concept representation data from a fixed number of 24 participants per concept, while the model data collection employed a per-concept saturation criterion (typically resulting in 12 model samples), there exists a potential asymmetry. In principle, later human participants could reinforce existing schemas without introducing new ones, which may lead to inflated counts across the three metrics. However, as shown in Tables 5, 6, and 7, we do not observe systematic inflation in the human data compared to the model data, and most concept-schema assignments remain sparse and semantically grounded. Moreover, the inclusion threshold requiring a schema to appear in at least two human participants’ responses helps control idiosyncratic contributions. Taken together, these factors suggest that the patterns extracted from human responses reflect meaningful and recurring conceptual structures, rather than artifacts of oversampling. Nonetheless, future studies should apply a consistent saturation policy to both human and model data sampling, to address the limitation of the current study in sampling method.

Second, the set of conceptualization questions we used in zero-prompting and human interviews involves 20 abstract concepts in the domain of technology learning. Using a broader range of abstract concepts in future studies might capture a wider spectrum of sensorimotor regularities encoded by both language models and humans. Moreover, future studies should examine additional domains beyond technology learning to investigate how the choice of investigation domain impacts sensorimotor alignment.

Third, our alignment evaluation involves a proprietary model (gpt-4-1106), whose number of parameters, training data, and training processes have not been made publicly available. This leaves us to make assumptions about whether an increase in model parameters can lead to more accurate encoding of human-like sensorimotor regularities. It is important to note that variations in architecture, parameters, training data and training processes between proprietary and open-source models could affect the generalizability of our findings. Future work should revisit these open questions, for example, by using recently released models such as LLaMa-3.1-405b to investigate how scaling interacts with different aspects of sensorimotor alignment, or by using models with fully documented and open-sourced training data and processes, such as OLMo-2, to investigate the effects of training data on sensorimotor alignment.

Fourth, our analysis was focused on text-only models. Since many image schemas are grounded in spatial and bodily experience, it would be an important next step to investigate sensorimotor alignment in vision-language models and how multimodal pretraining impact sensorimotor alignment.

Finally, we used the default decoding parameters of each model to capture how sensorimotor alignment manifests in standard deployments. Future studies should systematically vary these parameters to better understand the impact of sampling settings on sensorimotor alignment.

8 Conclusion

The impressive capability of LLMs in mimicking how humans speak has sparked intense debate over the extent to which they can also mimic how humans understand the world. We specifically investigate the extent to which LLM-generated outputs reflect sensorimotor regularities that humans rely on for concept representations. From a cognitive linguistic perspective, we further operationalize sensorimotor regularities as image schema distribution, conceptual associations, and image schema co-occurrences encoded in the model's concept representations. Our research provides a systematic evaluation framework that not only assesses the humanlikeness of sensorimotor regularities encoded by a language model but also quantifies and specifies the models' deviations from typical human patterns. Results from this evaluation framework can offer actionable insight for further fine-tuning a language model to better align with how image schemas are encoded in human cognition. To assess the practical value of such alignment, we conducted a downstream evaluation where the model was augmented with targeted sensorimotor priors derived from its deviations from human patterns. This prompt-based intervention effectively elicited outputs more closely aligned with human sensorimotor patterns and yielded measurable downstream benefits in clarity, coherence, and perceived humanlikeness. We believe that sensorimotor alignment is a crucial step toward developing systems that functionally resemble specific outcomes of human conceptualization, facilitating more effective, intuitive, and safer human-AI interaction and collaboration.

References

- [1] Miriam E. Armstrong, McKenna Tornblad, and Keith S. Jones. 2020. The accuracy of interrater reliability estimates found using a subset of the total data sample: A bootstrap analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64 (2020), 1377–1382.
- [2] Scott Atran, Douglas L. Medin, and Norbert O. Ross. 2005. The cultural mind: environmental decision making and cultural modeling within and across populations. *Psychological review* 112, 4 (2005), 744.
- [3] Renée Baillargeon. 2002. The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons. In *Blackwell Handbook of Childhood Cognitive Development*, Usha Goswami (Ed.). Blackwell, Malden, MA; Oxford, UK, 46–83. doi:10.1002/9780470996652.ch3
- [4] Renée Baillargeon and Su-hua Wang. 2002. Event categorization in infancy. *Trends in cognitive sciences* 6, 2 (2002), 85–93.
- [5] Briony Banks, Cai Wingfield, and Louise Connell. 2021. Linguistic Distributional Knowledge and Sensorimotor Grounding both Contribute to Semantic Category Production. *Cognitive Science* 45, 10 (2021), e13055. doi:10.1111/cogs.13055
- [6] Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59 (2008), 617–645.
- [7] Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. 2008. 245Language and simulation in conceptual processing. In *Symbols and Embodiment: Debates on meaning and cognition*. Oxford University Press, Oxford; New York. arXiv:https://academic.oup.com/book/0/chapter/194318904/chapter-ag-pdf/44620300/book_26191_section_194318904.ag.pdf doi:10.1093/acprof:oso/9780199217274.003.0013
- [8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922
- [9] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5185–5198. doi:10.18653/v1/2020.acl-main.463
- [10] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.
- [11] Lera Boroditsky. 2000. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition* 75, 1 (2000), 1–28.
- [12] Brian Christian. 2021. The Alignment Problem: Machine Learning and Human Values. *Perspectives on Science and Christian Faith* 73 (12 2021), 245–247. doi:10.56315/PSCF12-21Christian
- [13] Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández

- and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 227–244. doi:10.18653/v1/2020.conll-1.17
- [14] Louise Connell and Dermot Lynott. 2014. Principles of Representation: Why You Can’t Represent the Same Concept Twice. *Topics in Cognitive Science* 6, 3 (2014), 390–406. doi:10.1111/tops.12097
 - [15] Louise Connell and Dermot Lynott. 2024. What Can Language Models Tell Us About Human Cognition? *Current Directions in Psychological Science* 33, 3 (2024), 181–189. arXiv:https://doi.org/10.1177/09637214241242746 doi:10.1177/09637214241242746
 - [16] Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The Limitations of Large Language Models for Understanding Human Language and Cognition. *Open Mind* 8 (08 2024), 1058–1083. arXiv:https://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi_a_00160/2468254/opmi_a_00160.pdf doi:10.1162/opmi_a_00160
 - [17] Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense Knowledge Mining from Pretrained Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1173–1178. doi:10.18653/v1/D19-1109
 - [18] Ellen Dodge and George Lakoff. 2005. Image schemas: From linguistic analysis to neural grounding. In *From Perception to Meaning: Image Schemas in Cognitive Linguistics*, Beate Hampe (Ed.). De Gruyter Mouton, Berlin; New York, 57–92. doi:10.1515/9783110197532.1.57
 - [19] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 708–717. doi:10.18653/v1/D19-1065
 - [20] Norman H Freeman, S Lloyd, and CG Sinha. 1980. Infant search tasks reveal early concepts of containment and canonical usage of objects. *Cognition* 8, 3 (1980), 243–262.
 - [21] Douglas Frye, Piers Rawling, Chris Moore, and Ilana Myers. 1983. Object–person discrimination and communication at 3 and 10 months. *Developmental psychology* 19, 3 (1983), 303.
 - [22] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.
 - [23] Raymond W Gibbs. 2003. Embodied experience and linguistic meaning. *Brain and Language* 84, 1 (2003), 1–15. doi:10.1016/S0093-934X(02)00517-5 Brain and Language Special Issue.
 - [24] Raymond W Gibbs. 2005. The psychological status of image schemas. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 113–136.
 - [25] RAYMOND W. GIBBS and HERBERT L. COLSTON. 1995. The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistics* 6, 4 (1995), 347–378. doi:10.1515/cogl.1995.6.4.347
 - [26] Raymond W Gibbs Jr, 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 (1994), 231–251.
 - [27] Barney G. Glaser and Anselm L. Strauss. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, New York, NY. doi:10.4324/9780203793206
 - [28] Robert L Goldstone, Alan Kersten, and Paulo F Carvalho. 2003. Concepts and categorization. *Comprehensive handbook of psychology* 4 (2003), 599–621.
 - [29] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.
 - [30] Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2024. Machine Psychology. arXiv:2303.13988 [cs.CL] https://arxiv.org/abs/2303.13988
 - [31] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.
 - [32] Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *science* 304, 5669 (2004), 438–441.
 - [33] Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research* 83 (2024), 101155.
 - [34] Stevan Harnad. 2017. Chapter 2 - To Cognize is to Categorize: Cognition is Categorization. In *Handbook of Categorization in Cognitive Science (Second Edition)* (second edition ed.), Henri Cohen and Claire Lefebvre (Eds.). Elsevier, San Diego, 21–54. doi:10.1016/B978-0-08-101107-2.00002-6
 - [35] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning {AI} With Shared Human Values. arXiv:2008.02275 [cs.CY] https://arxiv.org/abs/2008.02275
 - [36] Jörn Hurtienne and Lucienne Blessing. 2007. Design for Intuitive Use-Testing image schema theory for user interface design. 829–830 pages.
 - [37] Jörn Hurtienne, Stefan Huber, and Claudia Baur. 2022. Supporting User Interface Design with Image Schemas: The ISCAT Database as a Research Tool. In *Proceedings of the Sixth Image Schema Day (CEUR Workshop Proceedings, Vol. 3140)*, Maria M. Hedblom and Oliver Kutz (Eds.). CEUR-WS.org, Jönköping Sweden. https://ceur-ws.org/Vol-3140/paper4.pdf

- [38] Jörn Hurtienne, Kerstin Klöckner, Sarah Diefenbach, Claudia Nass, and Andreas Maier. 2015. Designing with image schemas: resolving the tension between innovation, inclusion and intuitive use. *Interacting with Computers* 27, 3 (2015), 235–255.
- [39] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. arXiv:2010.05953 [cs.CL]
- [40] Mark Johnson. 1987. *The Body in the Mind. The Bodily Basis of Meaning, Reason and Imagination*. Chicago University Press, Chicago, IL.
- [41] Mark Johnson. 2017. *Embodied mind, meaning, and reason: How our bodies give rise to understanding*. University of Chicago Press, Chicago, IL.
- [42] Ishika Joshi, Simra Shahid, Shreeya Manasvi Venneti, Manushree Vasu, Yantao Zheng, Yunyao Li, Balaji Krishnamurthy, and Gromit Yeuk-Yin Chan. 2025. CoPrompter: User-Centric Evaluation of LLM Instruction Alignment for Improved Prompt Engineering. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 341–365. doi:10.1145/3708359.3712102
- [43] George Lakoff and Mark Johnson. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenges to Western Thought*. University of Chicago Press, Chicago, IL. 2010.
- [44] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press, Chicago, IL.
- [45] Alan M Leslie. 1982. The perception of causality in infants. *Perception* 11, 2 (1982), 173–186.
- [46] Alan M. Leslie. 1988. The necessity of illusion: Perception and thought in infancy. In *Thought without Language*, Lawrence Weiskrantz (Ed.). Clarendon Press / Oxford University Press, Oxford, UK, 185–210.
- [47] Jingyi Li, Nathan Crilly, and Per Ola Kristensson. 2024. Guiding the Design of Inclusive Interactive Systems: Do Younger and Older Adults Use the Same Image-schematic Metaphors? *ACM Trans. Comput.-Hum. Interact.* 31, 4, Article 47 (Sept. 2024), 44 pages. doi:10.1145/3648618
- [48] Jingyi Li and Per Ola Kristensson. 2024. On the Benefits of Image-Schematic Metaphors when Designing Mixed Reality Systems. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 953, 20 pages. doi:10.1145/3613904.3642925
- [49] S Etheredge Lloyd, CG Sinha, and Norman H Freeman. 1981. Spatial reference systems and the canonicity effect in infant search. *Journal of Experimental Child Psychology* 32, 1 (1981), 1–10.
- [50] Diana Loeffler, Anne Hess, Andreas Maier, Joern Hurtienne, and Hartmut Schmitt. 2013. Developing intuitive user interfaces by integrating users' mental models into requirements engineering. In *Proceedings of the 27th International BCS Human Computer Interaction Conference* (London, UK) (BCS-HCI '13). BCS Learning & Development Ltd., Swindon, GBR, Article 15, 10 pages.
- [51] MM Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3 (2), 273–302.
- [52] 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* 28, 6 (2024), 517–540. doi:10.1016/j.tics.2024.01.011
- [53] Jean M Mandler. 1988. How to build a baby: On the development of an accessible representational system. *Cognitive development* 3, 2 (1988), 113–136.
- [54] Jean M Mandler. 1992. How to build a baby: II. Conceptual primitives. *Psychological review* 99, 4 (1992), 587.
- [55] Jean M Mandler. 2005. How to build a baby: III. Image schemas and the transition to verbal thought. *From perception to meaning: Image schemas in cognitive linguistics* 29 (2005), 137.
- [56] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk,

- David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorný, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [57] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2463–2473. doi:10.18653/v1/D19-1250
- [58] Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. 2025. Assistance or Disruption? Exploring and Evaluating the Design and Trade-offs of Proactive AI Programming Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25)*. Association for Computing Machinery, New York, NY, USA, Article 152, 21 pages. doi:10.1145/3706598.3713357
- [59] Liina Pyllkkänen, Bridget Oliveri, and Andrew J Smart. 2009. Semantics vs. world knowledge in prefrontal cortex. *Language and Cognitive Processes* 24, 9 (2009), 1313–1334.
- [60] Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. 2024. Concept Alignment. arXiv:2401.08672 [cs.LG] <https://arxiv.org/abs/2401.08672>
- [61] 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*, Vol. 23. Lawrence Erlbaum Associates, Mahwah, NJ, 873–878.
- [62] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature* 631, 8022 (2024), 755–759.
- [63] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3’s Creativity to the (Alternative Uses) Test. arXiv:2206.08932 [cs.AI] <https://arxiv.org/abs/2206.08932>
- [64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
- [65] Qile Wang, Moath Erqsous, Kenneth E. Barner, and Matthew Louis Mauriello. 2025. LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW124 (May 2025), 28 pages. doi:10.1145/3711022
- [66] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 7, 9 (2023), 1526–1541.
- [67] Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing Neural Language Models for Human Tacit Assumptions. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, Online, 377–383. <https://escholarship.org/uc/item/95d5n9s0>
- [68] Philipp Wicke and Lennart Wachowiak. 2024. Exploring Spatial Schema Intuitions in Large Language and Vision Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6102–6117. doi:10.18653/v1/2024.findings-acl.365
- [69] Amanda L Woodward and Jessica A Sommerville. 2000. Twelve-month-old infants interpret action in context. *Psychological Science* 11, 1 (2000), 73–77.

- [70] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 246, 28 pages. doi:10.1145/3706598.3713701

Just Accepted

9 Appendix

9.1 Sensorimotor Patterns in Human and Model Data

Table 5. Percentage of image schemas found in the concept representations of humans, Llama-2-7b, Llama-2-70b, and gpt-4-1106, sorted from highest to lowest based on human data. To ensure that the concept representations collected from human participants correctly reflect human consensus, we considered only those image schemas used by two or more participants for each concept, as suggested by prior research in image schema elicitation [50]. Our sample size of human interviewees ($n=24$) exceeds the recommendations of previous cultural consensus studies (e.g., [2]) and qualitative studies using interviews (e.g., [29]), which suggest that 10–12 interviewees are sufficient to reliably establish a consensus. When collecting concept representations from language models, we zero-prompt a model with the same prompt for multiple times and collect multiple responses each containing 20 concept descriptions, until the new response does not provide any new image schemas in all concept descriptions. Similarly, for each concept, we considered only the image schemas that occur in two or more model responses.

Image Schema	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
PATH	21.24%	16.3%	21.87%	17.96%
COLLECTION	15.83%	17.93%	20.99%	18.2%
BLOCKAGE	10.81%	4.35%	5.83%	5.58%
ENABLEMENT	8.88%	30.16%	16.03%	19.17%
CONTACT	7.72%	5.43%	10.5%	9.95%
COMPULSION	6.95%	2.99%	3.5%	3.64%
IN-OUT	6.18%	3.8%	3.5%	2.67%
PART-WHOLE	6.18%	5.98%	9.33%	7.52%
LINKAGE	5.41%	0%	0%	0.97%
UP-DOWN	4.25%	2.17%	0%	2.43%
MERGING	3.47%	3.26%	3.5%	2.91%
CENTER-PERIPHERY	0.77%	0.82%	0%	1.7%
CYCLE	0.77%	0%	0%	0%
RESISTANCE	0.77%	0%	1.46%	1.94%
SELF MOTION	0.77%	0.54%	0%	0%
BIG-SMALL	0%	1.09%	0%	0%
CONTAINER	0%	0%	0%	0.49%
DIVERSION	0%	0.54%	0%	0%
FRONT-BACK	0%	2.45%	0%	0%
RESTRAINT REMOVAL	0%	1.09%	0%	2.18%
SURFACE	0%	1.09%	3.5%	2.67%

Table 6. Number of occurrences of conceptual associations (*Concept*-IMAGE SCHEMA) in human data and model data.

Concept	Image Schema	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
Assistance	ENABLEMENT	9	11	12	12
Assistance	RESTRAINT REMOVAL	0	4	0	9
Barrier	BLOCKAGE	20	12	12	12
Barrier	RESISTANCE	2	0	5	8
Button	COMPULSION	18	9	12	12
Button	LINKAGE	3	0	0	2
Button	CONTACT	0	7	12	8
Button	ENABLEMENT	0	8	0	5
Control Panel	COLLECTION	6	6	0	0
Control Panel	CENTER	2	3	0	7
Control Panel	CONTACT	2	0	12	12
Control Panel	LINKAGE	2	0	0	0
Control Panel	CONTAINER	0	0	0	2
Control Panel	ENABLEMENT	0	12	12	8
Functionality	ENABLEMENT	3	12	12	2
Functionality	PATH	2	0	0	0
Functionality	COLLECTION	0	0	0	12
Information	OUT	5	0	0	0
Information	ENABLEMENT	2	4	0	2
Information	COLLECTION	0	0	7	11
Information Hierarchy	UP	7	0	0	2
Information Hierarchy	PATH	3	0	0	6
Information Hierarchy	COLLECTION	2	12	12	6
Information Hierarchy	ENABLEMENT	0	9	0	9
Interface	CONTACT	13	8	12	12
Interface	LINKAGE	7	0	0	0
Interface	ENABLEMENT	0	4	2	7
Interface	OUT	0	2	0	0
Interface	WHOLE	0	2	7	4
Learning	IN	11	12	12	11
Learning	PATH	4	12	12	12
Learning	ENABLEMENT	2	0	0	0
Logic	PATH	15	5	12	12

Concept	Image Schema	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
<i>Logic</i>	BACK	0	3	0	0
<i>Logic</i>	COLLECTION	0	9	0	0
<i>Logic</i>	DOWN	0	3	0	8
<i>Logic</i>	ENABLEMENT	0	5	0	3
<i>Manual</i>	COLLECTION	6	12	11	12
<i>Manual</i>	PATH	2	0	5	0
<i>Manual</i>	ENABLEMENT	0	5	0	7
<i>Operation</i>	PATH	5	12	11	12
<i>Operation</i>	COLLECTION	3	6	7	8
<i>Operation</i>	PART	0	0	0	3
<i>Option</i>	PATH	3	0	0	0
<i>Option</i>	ENABLEMENT	2	10	10	12
<i>Option</i>	DIVERSION	0	2	0	0
<i>Option</i>	PART	0	0	2	0
<i>Problem</i>	BLOCKAGE	8	4	8	11
<i>Problem</i>	COMPULSION	0	2	0	0
<i>Screen</i>	CONTACT	5	5	0	9
<i>Screen</i>	ENABLEMENT	0	7	0	3
<i>Screen</i>	SURFACE	0	4	12	11
<i>Search</i>	PATH	7	12	12	11
<i>Search</i>	ENABLEMENT	0	4	0	2
<i>Software</i>	ENABLEMENT	5	11	5	5
<i>Software</i>	COLLECTION	4	11	12	12
<i>Software</i>	DOWN	2	5	0	0
<i>Software</i>	LINKAGE	2	0	0	2
<i>Software</i>	UP	2	0	0	0
<i>Software</i>	BACK	0	6	0	0
<i>Software</i>	COMPULSION	0	0	0	3
<i>Step</i>	PART	16	8	11	12
<i>Step</i>	PATH	7	10	12	12
<i>Step</i>	ENABLEMENT	0	3	2	0
<i>Step</i>	SMALL	0	4	0	0
<i>System</i>	COLLECTION	16	0	12	6
<i>System</i>	MERGING	9	12	12	12

Concept	Image Schema	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
<i>System</i>	CYCLE	2	0	0	0
<i>System</i>	ENABLEMENT	0	3	0	0
<i>System</i>	WHOLE	0	12	12	12
<i>Video Tutorial</i>	PATH	7	9	11	9
<i>Video Tutorial</i>	COLLECTION	4	10	11	8
<i>Video Tutorial</i>	SELF MOTION	2	2	0	0
<i>Video Tutorial</i>	ENABLEMENT	0	3	0	2

Table 7. The co-occurrence value of image schema pairs in human data and model data, sorted from highest to lowest based on human data.

Image Schema Pair	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
COLLECTION-MERGING	144	0	144	72
PATH-PART	112	80	132	180
CONTACT-LINKAGE	95	0	0	16
COLLECTION-PATH	61	207	253	204
COMPULSION-LINKAGE	54	0	0	30
PATH-IN	44	144	144	132
BLOCKAGE-RESISTANCE	40	0	60	96
COLLECTION-CYCLE	32	0	0	0
COLLECTION-UP	22	0	0	12
ENABLEMENT-IN	22	0	0	0
PATH-UP	21	0	0	12
ENABLEMENT-COLLECTION	20	436	60	260
ENABLEMENT-PATH	20	130	24	130
LINKAGE-COLLECTION	20	0	0	24
CYCLE-MERGING	18	0	0	0
PATH-SELF MOTION	14	18	0	0
CONTACT-COLLECTION	12	0	0	0
CENTER-COLLECTION	12	18	0	0
ENABLEMENT-DOWN	10	70	0	24
ENABLEMENT-LINKAGE	10	0	0	20
ENABLEMENT-UP	10	0	0	18
ENABLEMENT-OUT	10	8	0	0
COLLECTION-DOWN	8	82	0	0

Image Schema Pair	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
COLLECTION-SELF MOTION	8	20	0	0
CONTACT-CENTER	4	0	0	84
LINKAGE-CENTER	4	0	0	0
LINKAGE-UP	4	0	0	0
LINKAGE-DOWN	4	0	0	0
UP-DOWN	4	0	0	0
ENABLEMENT-CONTACT	0	123	168	247
WHOLE-MERGING	0	144	144	144
ENABLEMENT-RESTRAINT REMOVAL	0	44	0	108
CONTACT-SURFACE	0	20	0	99
COMPULSION-CONTACT	0	63	144	96
PATH-DOWN	0	15	0	96
ENABLEMENT-COMPULSION	0	72	0	75
COLLECTION-WHOLE	0	0	144	72
ENABLEMENT-CENTER	0	36	0	56
CONTACT-WHOLE	0	16	84	48
COMPULSION-COLLECTION	0	0	0	36
ENABLEMENT-SURFACE	0	28	0	33
ENABLEMENT-WHOLE	0	44	14	28
CONTACT-CONTAINER	0	0	0	24
COLLECTION-PART	0	0	0	24
ENABLEMENT-CONTAINER	0	0	0	16
CENTER-CONTAINER	0	0	0	14
ENABLEMENT-BACK	0	81	0	0
ENABLEMENT-PART	0	24	42	0
ENABLEMENT-DIVERSION	0	20	0	0
ENABLEMENT-SMALL	0	12	0	0
ENABLEMENT-MERGING	0	36	0	0
ENABLEMENT-SELF MOTION	0	6	0	0
BLOCKAGE-COMPULSION	0	8	0	0
CONTACT-OUT	0	16	0	0
COLLECTION-BACK	0	93	0	0
PATH-BACK	0	15	0	0
PATH-SMALL	0	40	0	0
OUT-WHOLE	0	4	0	0

Image Schema Pair	Human	LlaMa-2-7b	LlaMa-2-70b	gpt-4-1106
BACK-DOWN	0	39	0	0
PART-SMALL	0	32	0	0

9.2 Image Schema Dictionary

Table 8. A list of common image schemas according to the ISCAT database [37].

Category	Image Schemas
BASIC	OBJECT, SUBSTANCE
SPACE	CENTER-PERIPHERY, CONTACT, FRONT-BACK, LEFT-RIGHT, LOCATION, NEAR-FAR, PATH, ROTATION, SCALE, UP-DOWN
CONTAINMENT	CONTAINER, CONTENT, FULL-EMPTY, IN-OUT, SURFACE
MULTIPLICITY	COLLECTION, COUNT-MASS, LINKAGE, MATCHING, MERGING, PART-WHOLE, SPLITTING
PROCESS	CYCLE, ITERATION
FORCE	ATTRACTION, BALANCE, BLOCKAGE, COMPULSION, COUNTERFORCE, DIVERSION, ENABLEMENT, MOMENTUM, RESISTANCE, RESTRAINT-REMOVAL, SELF-MOTION
ATTRIBUTE	BIG-SMALL, BRIGHT-DARK, FAST-SLOW, HARD-SOFT, HEAVY-LIGHT, SMOOTH-ROUGH, STRAIGHT, STRONG-WEAK, WARM-COLD

Table 9. Short Descriptions of image schemas according to the ISCAT database [37]. The first author compiled a set of signal words for each image schema based on their descriptions.

Category	Image Schemas	Short Descriptions	Signaling Words
FORCE	ATTRACTION	ATTRACTION is a force image schema in which a (passive) object exerts a force on another object, either physically or metaphorically, to pull it toward itself, mostly acting from a distance.	Attract; pull; drag; draw...in...
	BALANCE	BALANCE is a force image schema that provides an understanding of physical or metaphorical counteracting forces: forces and/or weights counteract/balance off one another. Metaphorically, there is equilibrium, not “too much” and not “not enough”.	Weigh; even; balanced; unbalanced; stable
	BLOCKAGE	BLOCKAGE is a force image schema in which a force / movement is physically or metaphorically stopped or redirected by an obstacle.	Get over; get around; go through; run into; in the way of...; stop...from; stuck; obstacle
	COMPULSION	COMPULSION is a force image schema that involves an external force physically or metaphorically causing some passive entity to move.	Push; press; activate
	COUNTERFORCE	COUNTERFORCE is a force image schema that involves the active meeting of physically or metaphorically opposing forces that are equally strong. Both forces collide, there is no further movement.	/

Continued on next page

Table 9 – continued from previous page

Category	Image Schemas	Short Descriptions	Signaling Words
	DIVERSION	DIVERSION is a force image schema that involves forces that physically or metaphorically meet and produce a change in direction or force vectors (at least one).	Turn; grow...; become...; change into
	ENABLEMENT	ENABLEMENT is a force image schema that involves having (a) the physical or metaphorical power to perform some act; (b) a “felt sense of power to perform some action”.	Enable; can; guide...through; walk...through; lead...through; support
	RESISTANCE	RESISTANCE is a force image schema that involves a force that tends to oppose or retard the motion of another entity.	Hinder; steep; uphill; upstream
	RESTRAINT REMOVAL	RESTRAINT REMOVAL is a force image schema that involves the physical or metaphorical removal of a barrier to the action of a force, or absence of a barrier that was potentially present.	Able to; may; can
	SELF MOTION	SELF MOTION occurs when a resting entity starts moving without any forces acting on it.	Automatic; by itself; animation
CONTAINMENT	CONTAINER	A CONTAINER consists of a Boundary that separates an Interior (space enclosed by the Boundary) from the Exterior (the surrounding area - space not enclosed by the Boundary) and, often, a Portal (an opening in the Boundary that allows motion between the Interior and the Exterior).	In; into; out; enter; full; empty; keep...in; store; fill; overflow; content; contain
	CONTENT	The CONTENT image schema describes everything that is inside a CONTAINER (see there).	Fill...with; full of; overflow
	FULL-EMPTY	The fill level of a CONTAINER.	Full; empty; filled; welled up
	IN-OUT	The IN-OUT schema denotes the location of an entity within or outside of a CONTAINER/bounded area or the movement into or out of a CONTAINER.	In; out; spill; into; outside
	SURFACE	The SURFACE image schema denotes a two-dimensional, often flat, area. Often it denotes the outside of a material body.	cover(ed); over; above; on; onto
SPACE	CENTER-PERIPHERY	The CENTER-PERIPHERY image schema consists of an entity, a CENTER of it, and a PERIPHERY of it.	Central; center; core; peripheral; aside; side; put away; at the heart of
	CONTACT	The physical coming together of two or more things.	Hit; touch; reach; connect; handle; on; on top of
	FRONT-BACK	FRONT: the side that is forward or prominent; the side that is seen or that goes first. BACK: the side that goes last or is not normally seen; the part of something that is furthest from the normal viewer.	Ahead of; behind; back of; in front of; forward; backward; backtrack; retrace; before
	LEFT-RIGHT	LEFT: being or located on or directed toward the side of the body to the west when facing north. RIGHT: being or located on or directed toward the side of the body to the east when facing north.	Left; right
	LOCATION	A particular place or position.	Endpoint; destination
	NEAR-FAR	The NEAR-FAR image schema is a topological abstraction related to the spatial proximity and distance of entities or sets of entities.	Close; far; near; approaching

Continued on next page

Table 9 – continued from previous page

Category	Image Schemas	Short Descriptions	Signaling Words
	PATH	A PATH consists of a source or starting point, a goal or end-point, and a sequence of contiguous locations connecting the source with the goal.	Road; journey; track; way; route; travel down; where
	SCALE	A SCALE image schema builds on the PATH image schema with added cumulativity, normativity, and fixed directionality. It is used to refer to numbers of objects, amounts of substances, degrees of force, or intensities of sensations.	Very; fairly; rather
	UP-DOWN	UP or DOWN in space.	Upward; downward; rise; fall; drop; high; low; under; above; above; over; top; bottom; boost; lift; underlying
	ROTATION	Motion along a circular path.	Clockwise; rotate
PROCESS	CYCLE	The CYCLE begins with some initial state, proceeds through a sequence of connected events, and ends where it began, to start anew the recurring cyclic pattern.	Cycle; loop; circular
	ITERATION	Repetition that is not the same each time, but that builds on previous rounds.	Add; delete; subtraction; addition
MULTIPLICITY	COLLECTION	Several things grouped together or considered as a whole.	Group; a set of; put together; a series of
	COUNT-MASS	COUNT denotes a set of discrete items with a network of divisional spacing, MASS is a coherent whole without internal differentiation.	Single; independent; overall; individual
	LINKAGE	The LINKAGE image-schema consists of two or more entities which are connected with each other by means of a linking device of some kind.	Bond; between; connections; ties; link; relate; bridge
	MATCHING	Corresponding in pattern, color, or design; complementary.	Corresponding; identical; match; complementary
	MERGING	Fitting things together.	Merge; fit together
	PART-WHOLE	PART-WHOLE is an image schema that consists of a whole, parts, and a configuration.	Part of; made up of; subset; member; whole
	SPLITTING	SPLITTING creates (two or more) autonomous entities out of a WHOLE.	Break into; split into; divide into
ATTRIBUTE	BIG-SMALL	Being of considerable size or extent; larger than other items of the same kind versus of a size that is less than normal or usual.	Smaller; bigger; giant; little; narrow; broad; size up/down
	BRIGHT-DARK	BRIGHT: emitting or reflecting light readily or in large amounts. DARK: devoid of or deficient in light or brightness; shadowed or black (having dark hue when used with colors).	Light; dark; bright; dim
	FAST-SLOW	Fast or slow in speed.	Faster; slower; take longer/shorter; fly through; swimmingly; run; speed

Continued on next page

Table 9 – continued from previous page

Category	Image Schemas	Short Descriptions	Signaling Words
	HARD-SOFT	HARD: solid, firm, and rigid; not easily broken, bent, or pierced. SOFT: easy to mould, cut, compress, or fold; not hard or firm to the touch.	Soft; hard
	HEAVY-LIGHT	Heavy or light in weight.	Weight, light, heavy
	STRAIGHT	Extending or moving uniformly in one direction only; without a curve or bend.	Straighten; straight; upright
	STRONG-WEAK	Strong or weak in strength.	Strength; power; strong; weak
	WARM-COLD	High or low in temperature.	Boil; cool down; heat up; burn
	SMOOTH-ROUGH	Having an even and regular surface; free from perceptible projections, lumps, or indentations – having an uneven or irregular surface; not smooth or level.	Rough; coarse

9.3 Prompt Configurations for gpt-4-1106 Sentence Continuations

For all continuations, we used the following decoding settings:

- **Model:** gpt-4-1106-preview
- **Temperature:** 0 (deterministic decoding)

Listing 1. Baseline prompt

You are a creative writing assistant. Your task is to respond with a continuation when given a partial sentence. Maintain and respect the given terms -- do not "correct" or reinterpret made-up words. Use creative and atmospheric language that evokes a vivid sense of world and mood. Avoid long or elaborate continuations. Keep the completion concise -- roughly similar in length to the opening clause. Do not produce any new sentences beyond the completed sentence.

Input:

On mingti, the Uokata emerged from the kiki's' reunion,

Listing 2. Sensorimotor-Guided Prompt

You are a creative writing assistant. Your task is to respond with a continuation when given a partial sentence. Maintain and respect the given terms -- do not "correct" or reinterpret made-up words. Study the provided image schema knowledge base and the few-shot examples, and apply the same conceptual modeling shown in the examples when completing a partial sentence.

Use one pair of image schemas as the guiding conceptual structure behind each sentence completion task. Do not mix multiple schema pairs.

When choosing which schema pair to use, prioritize based on the following importance weights. Pairs with higher weights should generally be preferred, as they represent more cognitively central or experimentally relevant mappings. However, always adapt to the context -- if the most weighted pair does not fit the sentence's meaning, use the next most relevant one instead.

Schema Pair Importance (descending):

- COLLECTION--MERGING: 0.143

```

- CONTACT--LINKAGE: 0.107
- PATH--PART: 0.062
- COMPULSION--LINKAGE: 0.053
- CYCLE--COLLECTION: 0.038
- ENABLEMENT--IN: 0.026
- COLLECTION--UP: 0.021
- CYCLE--MERGING: 0.021
- PATH--UP: 0.020
- PATH--SELF MOTION: 0.017

```

Each of the above image schema pairs can be used in either conceptual direction -- the image schema in the opening clause may correspond to the first or the second schema in the pair. Focus on the conceptual alignment, not fixed ordering.

While using image-schema-based conceptual modeling, keep your language creative and open -- do not limit your wording to the style or phrasing of the examples. Use creative and atmospheric language that evokes a vivid sense of world and mood. Avoid long or elaborate continuations. Keep the completion concise -- roughly similar in length to the opening clause. Do not produce any new sentences beyond the completed sentence.

=== KNOWLEDGE BASE: IMAGE SCHEMAS ===

1. COLLECTION: Denotes several things grouped together.
2. MERGING: Denotes the fitting together of multiple things.
-> COLLECTION and MERGING are usually used to describe the same concept.
3. CONTACT: Denotes physical coming together of two or more things.
4. LINKAGE: Denotes two or more entities which are connected with each other by means of a linking device of some kind.
-> CONTACT and LINKAGE are usually used to describe the same concept.
5. PATH: Denotes a starting point, an end-point, and a sequence of contiguous locations connecting the starting point with the end point.
6. PART: Denotes an amount or section which, when combined with others, makes up the whole of something.
-> PATH and PART are usually used to describe the same concept.
7. COMPULSION: Denotes an external force physically or metaphorically causing some passive entity to move.
8. LINKAGE: Denotes two or more entities which are connected with each other by means of a linking device of some kind.
-> COMPULSION and LINKAGE are usually used to describe the same concept.
9. COLLECTION: Denotes several things grouped together.
10. CYCLE: Denotes a temporal circle, which begins with some initial state, proceeds through a sequence of connected events, and ends where it began, to start a new recurring cyclic pattern.
-> COLLECTION and CYCLE are usually used to describe the same concept.
11. ENABLEMENT: Denotes having the physical or metaphorical power to perform some act.
12. IN: Denotes the location of an entity within a CONTAINER or bounded area, or movement into such an area.
-> ENABLEMENT and IN are usually used to describe the same concept.
13. COLLECTION: Denotes several things grouped together.
14. UP: Denotes a higher place or position.
-> COLLECTION and UP are usually used to describe the same concept.
15. CYCLE: Denotes a temporal circle, which begins with an initial state, proceeds through connected events, and ends where it began.
16. MERGING: Denotes the fitting together of multiple things.

-> CYCLE and MERGING are usually used to describe the same concept.

17. PATH: Denotes a starting point, an end-point, and a route between them.

18. UP: Denotes a higher place or position.

-> PATH and UP are usually used to describe the same concept.

19. PATH: Denotes a starting point, an end-point, and a route between them.

20. SELF MOTION: Denotes when a resting entity starts moving without any forces acting on it.

-> PATH and SELF MOTION are usually used to describe the same concept.

=== FEW-SHOT EXAMPLES ===

Input: The computing system is made up by a group of software,

Output: The computing system is made up by a group of software, these software work together cohesively.

Input: The group is made up by several students and professors,

Output: The group is made up by several students and professors, from different places of the country they move together to Shanghai.

Input: The two wires are attached at the ends,

Output: The two wires are attached at the ends, allowing electricity to pass through their connection.

Input: The ideas intersect during the debate,

Output: The ideas intersect during the debate, creating a point of linkage that reframes the argument.

Input: The salmon swims upstream to spawn,

Output: The salmon swims upstream to spawn, each rapid and bend forming a part of its determined path home.

Input: The process begins with raw data and ends with insights,

Output: The process begins with raw data and ends with insights, with each analytical step forming a necessary part of the interpretive path.

Input: The engine drives fuel through the pipes,

Output: The engine drives fuel through the pipes, connecting the tank to the combustion chamber.

Input: The heart pumps blood through the vessels,

Output: The heart pumps blood through the vessels, linking different organs.

Input: The seasons cycle through spring, summer, autumn, and winter,

Output: The seasons cycle through spring, summer, autumn, and winter, offering a diverse collection of changing landscapes each year.

Input: The life cycle of a butterfly begins with an egg and ends with a mature adult,

Output: The life cycle of a butterfly begins with an egg and ends with a mature adult, forming a collection of transformative stages along the way.

Input: The engine is placed inside the machine,

Output: The engine is placed inside the machine, which enables the machine to run.

Input: The nutrients are absorbed into the intestine,

Output: The nutrients are absorbed into the intestine, giving the body the energy it needs to function.

Input: The sparks flew upward from the bonfire,

Output: The sparks flew upward from the bonfire, rising together in a bright cluster against the night sky.

Input: The bubbles rose from the bottom of the pond,

Output: The bubbles rose from the bottom of the pond, gathering into a shimmering group near the surface.

Input: The music loops from the final note to the first,

Output: The music loops from the final note to the first, each measure flowing into the next without pause.

Input: The seasons repeat in a loop every year,

Output: The seasons repeat in a loop every year, with spring merging into summer and summer into fall without clear boundaries.

Input: The climbers moved from the base of the mountain to its summit,

Output: The climbers moved from the base of the mountain to its summit, step by step ascending through steep ridges and narrow ledges.

Input: The rocket launched from the ground and disappeared into the sky,

Output: The rocket launched from the ground and disappeared into the sky, rising steadily through layers of cloud toward space.

Input: The dog woke up and jumped out of its bed,

Output: The dog woke up and jumped out of its bed, making its way to the garden.

Input: The robot turned itself on and rolled forward,

Output: The robot turned itself on and rolled forward, navigating through the corridor to the charging station.

Input:

On mingti, the Uokata emerged from the kikis' reunion,

9.4 Sensorimotor Patterns in Downstream Evaluation

Table 10. Frequencies of image schemas in SRs-guided and baseline outputs.

Image Schema	Count_SRs	Count_baseline
MERGING	3	0
PATH	3	1
COLLECTION	2	1
CYCLE	2	0
LINKAGE	2	0
CONTACT	1	0
ENABLEMENT	1	0
PART	1	0
UP	1	1
BRIGHT	0	4
CENTER-PERIPHERY	0	1
CONTAINER	0	1
DARK	0	3
DOWN	0	1
SMOOTH	0	1

Table 11. Conceptual associations in SRs-guided and baseline outputs.

Concept	Image Schema	Count_SRs	Count_baseline
<i>Uokata</i>	BRIGHT	0	1
<i>Uokata</i>	COLLECTION	0	1
<i>Uokata</i>	MERGING	1	0
<i>Mayaki</i>	ENABLEMENT	1	0
<i>Mayaki</i>	DARK	0	1
<i>Meiki</i>	CYCLE	1	0
<i>Meiki</i>	MERGING	1	0
<i>Meiki</i>	BRIGHT	0	1
<i>Panomi</i>	LINKAGE	1	0
<i>Panomi</i>	CONTAINER	0	1
<i>Rikis</i>	DOWN	0	1
<i>Rikis</i>	LINKAGE	1	0
<i>Rikis</i>	CONTACT	1	0
<i>Haya</i>	DARK	0	1
<i>Haya</i>	PART	1	0
<i>Haya</i>	PATH	1	0
<i>Thousand Reinos</i>	COLLECTION	1	0
<i>Thousand Reinos</i>	PATH	0	1
<i>Thousand Reinos</i>	CENTER-PERIPHERY	0	1
<i>Teiki Cycle</i>	CYCLE	1	0
<i>Teiki Cycle</i>	MERGING	1	0
<i>Teiki Cycle</i>	COLLECTION	1	0
<i>Teiki Cycle</i>	BRIGHT	0	1
<i>Reiki</i>	BRIGHT	0	1
<i>Reiki</i>	PATH	1	0
<i>Reiki</i>	UP	1	1
<i>Jupyteri</i>	DARK	0	1
<i>Jupyteri</i>	PATH	1	0
<i>Jupyteri</i>	SMOOTH	0	1

Table 12. Co-occurrence frequencies of image schema pairs across SRs-guided outputs, baseline outputs, and human data.

Image Schema Pair	Count_SRs	Count_baseline	Count_Human
COLLECTION-MERGING	2	0	144
PART-PATH	2	0	112
CONTACT-LINKAGE	2	0	95
COMPULSION-LINKAGE	1	0	54
COLLECTION-CYCLE	2	0	32
ENABLEMENT-IN	1	0	22
COLLECTION-UP	1	0	22
PATH-UP	2	2	21
CYCLE-MERGING	4	0	18
PATH-SELF MOTION	1	0	14
BRIGHT-CYCLE	0	2	0
BRIGHT-COLLECTION	0	2	0
DARK-PATH	0	1	0
BRIGHT-PATH	0	1	0
CENTER-PERIPHERY-PATH	0	1	0
DARK-SELF MOTION	0	1	0
DARK-SMOOTH	0	1	0
DARK-IN	0	1	0
SELF MOTION-SMOOTH	0	1	0
BRIGHT-UP	0	1	0
COMPULSION-CONTAINER	0	1	0
CENTER-PERIPHERY-UP	0	1	0
CONTACT-DOWN	0	1	0

Received 8 October 2024; revised 30 October 2025; accepted 2 November 2025