nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02203-8

Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts

Received: 28 November 2023

Accepted: 31 March 2025

Published online: 04 June 2025

Check for updates

Qihui Xu[1,7] ✉, Yingying Peng[2,7], Samuel A. Nastase[3], Martin Chodorow[4,5], Minghua Wu[2] & Ping Li[2,6] ✉

To what extent can language give rise to complex conceptual representation? Is multisensory experience essential? Recent large language models (LLMs) challenge the necessity of grounding for concept formation: whether LLMs without grounding nevertheless exhibit human-like representations. Here we compare multidimensional representations of ~4,442 lexical concepts between humans (the Glasgow Norms[1], $N = 829$; and the Lancaster Norms[2], $N = 3,500$) and state-of-the-art LLMs with and without visual learning, across non-sensorimotor, sensory and motor domains. We found that (1) the similarity between model and human representations decreases from non-sensorimotor to sensory domains and is minimal in motor domains, indicating a systematic divergence, and (2) models with visual learning exhibit enhanced similarity with human representations in visual-related dimensions. These results highlight the potential limitations of language in isolation for LLMs and that the integration of diverse modalities can potentially enhance alignment with human conceptual representation.

Imagine learning about the concept of 'flower' without ever smelling a rose, touching the petals of a daisy or walking through a field of wildflowers. Can we truly represent the concept 'flower' in all its richness without sensorimotor experiences? This question invokes a longstanding debate about the interplay between physical experience and conceptual representation. On the one hand, theories of grounded cognition posit that our senses are our gateways to knowledge[3]; the physical experience of 'flowers' is integral to how we represent and process them. On the other hand, research with disembodied artificial neural network models[4–6] and congenitally blind and partially sighted people[7–10] show that learners can form conceptually rich representations from language alone, independent of direct sensory experience. For example, studies show that individuals born with limited vision can represent and respond to colour concepts similarly to those who can see[8–10]. When sensorimotor input is absent, to what extent can language alone inform our conceptual representation of the world? How indispensable is bodily experience in shaping our conceptual world?

Disentangling the various sources for conceptual formation is challenging. Although studies involving artificial models or blind and partially sighted people have provided valuable insights, they have several limitations. First, they often overlook the multidimensional nature of conceptual representation. Our representation of concepts is extensive and complex, encompassing areas not directly tied to sensorimotor experiences, such as emotional arousal and valence linked to the concept, as well as the direct sensations and actions encountered in connection with it[11,12]. For instance, processing the concept of

[1]Department of Psychology, Ohio State University, Columbus, OH, USA. [2]Department of Chinese and Bilingual Studies, Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong SAR, China. [3]Department of Psychology and the Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [4]Department of Psychology, Hunter College, City University of New York, New York, NY, USA. [5]Department of Psychology, Graduate Center, City University of New York, New York, NY, USA. [6]The PolyU-Hangzhou Technology and Innovation Research Institute, Hangzhou, China. [7]These authors contributed equally: Qihui Xu, Yingying Peng. ✉e-mail: xu.5430@osu.edu; ping2.li@polyu.edu.hk

'flower' may evoke not only the object 'flower' itself but also the visual perceptions of colours and shapes, the actions of touching the flower by hand and smelling it with the nose, its associated scents, textures, emotions and memories. Second, the limited scope of words tested, such as colour words (for example, ref. 9) or object words (for example, ref. 5) only, restricts external validity, failing to capture the breadth of concepts encountered in daily life, which encompasses not only objects and colour words but also action verbs, abstract concepts and more[10,13,14]. Moreover, there can be potential knowledge transfer across domains, which poses challenges for human-participant research in achieving rigorous control over the diverse domains of resources that may contribute to conceptual representation. Even without visual input, individuals can tap into other sensory channels such as touch and internal sensations, which have been shown to correlate with visual knowledge[2]. Therefore, to better distinguish between language-derived and sensorimotor-derived sources, it is crucial to consider a broad range of concept words that span a wide and systematic spectrum of conceptual representations (from non-sensorimotor aspects to sensory and motor aspects).

Recent advances in large language models (LLMs) offer a unique avenue to test the extent to which language alone can give rise to complex concepts[15–17]. LLMs have enabled us to (1) estimate what kinds of structure (and how much) can ultimately be extracted from large volumes of language alone[18–20] and (2) examine how different input modalities (for example, text versus images) influence learning processes[15,16]. Current LLMs have been trained on massive amounts of data, either constrained to the language domain (that is, large-scale text data as in GPT-3.5 and PaLM) or incorporating language and visual input (for example, GPT-4 and Gemini). Despite these limited input modalities, these models exhibit remarkably human-like performance in various cognitive tasks[6,21–23]. In the same way that LLMs demonstrate the feasibility of learning syntactic structure from surface-level language exposure alone[24,25], they may also have the capability of learning physical, grounded features of the world from language alone[26–28]. For example, some have argued that language itself can act as a surrogate 'body' for these models, reminiscent of the largely conceptualized and ungrounded colour knowledge in blind and partially sighted individuals[4,6]. This perspective aligns with previous research emphasizing the important role of language in providing rich cognitive and perceptual resources[29,30]. By contrast, others believe that multimodal experiences are essential for both humans and artificial models to grasp concepts more efficiently[16,31,32]. Unlike current LLMs such as GPT-3, which rely on vast amounts of text[20]—equivalent to 20,000 years of human reading[33]—real-world, interactive experiences may offer richer, more interconnected conceptual representations that facilitate knowledge transfer across domains, potentially reducing the need for such extensive linguistic input in model training.

The above theoretical debates motivate us in this study to investigate two research questions. The first research question is which aspects of human conceptual representation can be recovered by ungrounded state-of-the-art LLMs and which cannot. To address this question, we compared similarity of representations across ~4,442 word concepts between humans and two state-of-the-art LLM families (Fig. 1a) from OpenAI (GPT-3.5 and GPT-4) and Google (PaLM and Gemini), across a range of dimensions, spanning non-sensorimotor, sensory and motor domains. The domains were based on categories established in refs. 1,2, where each domain consists of several dimensions (see Table 1 for definitions for each dimension). These dimensions provide comprehensive coverage for understanding the spectrum of human lexical–conceptual processing explored in previous studies (for example, refs. 34–36), from socio-emotional aspects and abstract mental imagery, to direct bodily experience (Fig. 1c). Importantly, our classification into 'non-sensorimotor' and 'sensorimotor' domains is based on whether the measures directly assess specific sensorimotor experiences. This operational distinction does not imply that the 'non-sensorimotor' dimensions are devoid of embodiment—a topic that remains widely explored in current literature (for example, refs. 37,38). For example, questions that explicitly ask about specific sensory or motor experiences—such as how one would experience the concept 'flower' through smelling—are categorized under sensorimotor domains because they directly tap into modality-specific (that is, the typical senses) and effector-specific (that is, the typical action effectors) bodily experiences. By contrast, questions that do not directly specify particular sensorimotor experiences—such as emotional arousal about the concept 'flower'—are categorized as non-sensorimotor, though they could relate to bodily states (see the close relation between emotion and interoception in ref. 36). Despite this potential connection, previous literature shows that compared with non-sensorimotor dimensions such as concreteness, specific sensory ratings are more effective in facilitating lexical semantic processing[39]. The second research question is the potential value of additional visual inputs for concept formation in LLMs trained with both language and visual input modalities, compared with those trained solely within the language domain. To address this research question, we analysed whether the additional visual inputs provided to LLMs yield stronger alignment with humans on visual-related dimensions. Finally, we validated the response of the LLMs to ensure the validity of our results (see Supplementary Information, section 7.3, for discussion on the importance of validation). We report two key findings: (1) the similarity between LLMs and human representations decreased from non-sensorimotor to sensory domains and was minimal in motor domains, and (2) the models incorporating visual inputs exhibited enhanced similarity with human representations in vision-related dimensions. These findings suggest that learning solely within the language domain substantially recovers non-sensorimotor aspects of conceptual representations yet remains impoverished in sensorimotor aspects, particularly along motor dimensions. Furthermore, extending experience into the visual domain is associated with LLM's improved alignment with human representations in both visual and related dimensions such as imageability and haptic features, suggesting potential knowledge transfer through multimodal integration.

Results

We collected conceptual word ratings from LLMs (that is, ChatGPT models: GPT-3.5 and GPT-4; Google LLMs: PaLM and Gemini) and compared them with ratings generated by humans from the Glasgow[1] and Lancaster Norms[2] (see Methods for further details and Supplementary Information, section 7.1, on how word rating tasks capture key aspects of conceptual representation). The model prompt and design (Fig. 1b) for LLMs was standardized to match the instructions given to human participants, maintaining consistency with human-participant data collection. Each LLM was separately run for four rounds to ensure reliability (see Supplementary Information, section 1, for the agreement between these rounds).

Two common practices for measuring similarity were used to evaluate LLM versus human similarity - dimension-wise correlations and representational similarity analysis (RSA) ('Model–human alignment varies across domains' section). These measures enable the evaluation of LLM–human similarity from several angles: the strength with which a lexical concept is rated within each individual dimension, and how different lexical concepts are geometrically organized across dimensions. Next, we explored if the additional visual inputs provided to LLMs predict their alignment with humans ('Linking additional visual training to model–human alignment' section). We then performed two secondary analyses to substantiate the correlations found between human and model ratings ('Validation of the results' section). Given ongoing debates regarding the distinct roles of grounding in concrete versus abstract concepts[40,41], we assessed the potential influence of word concreteness on our primary findings. To ensure the validity of LLMs as cognitive models[42], we adopted standard validation techniques from

https://doi.org/10.1038/s41562-025-02203-8

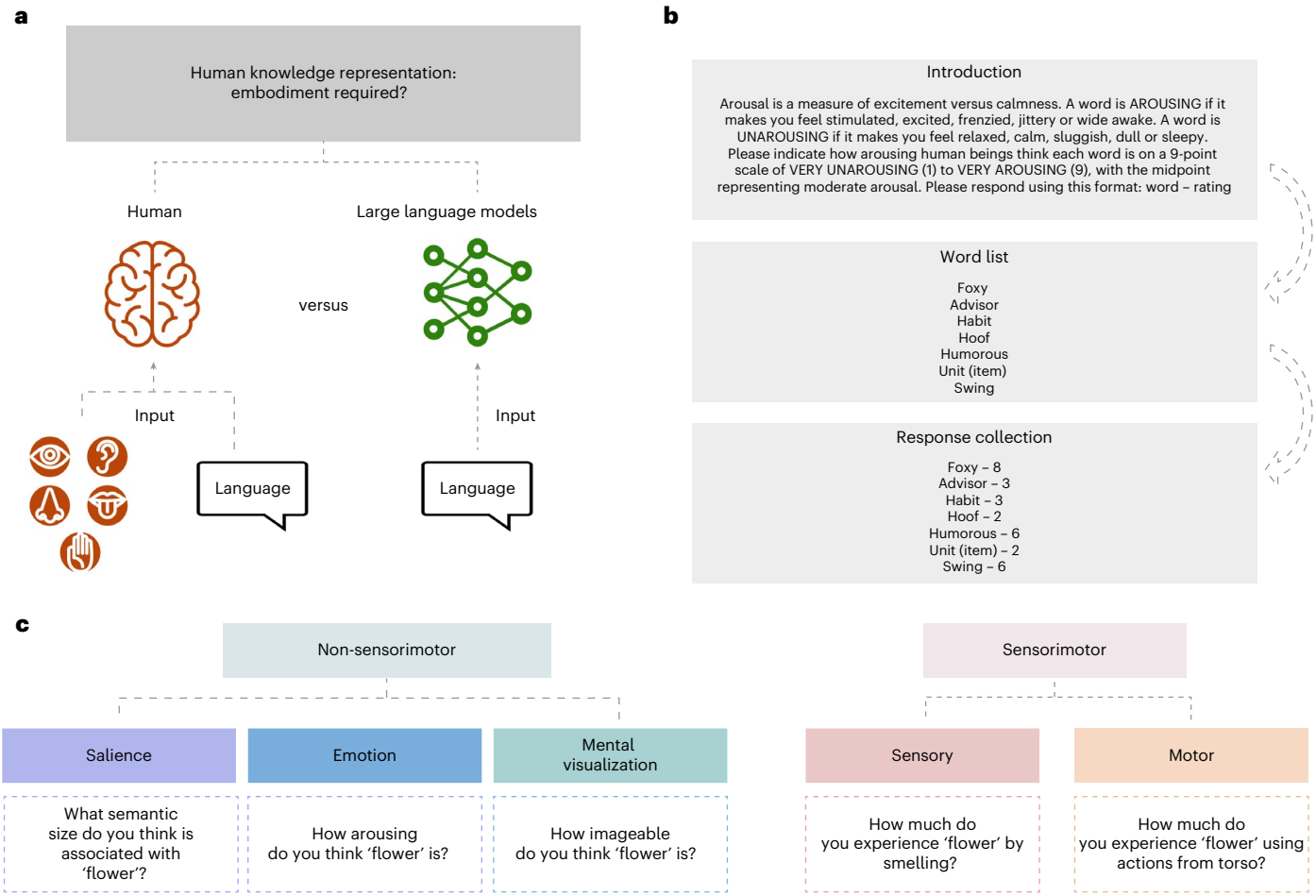

Fig. 1 | Overview. a, A schematic depiction of the research question and approach. This study aims to investigate the extent to which human conceptual representation requires grounding. Icons from Flaticon.com[77]. **b**, A schematic of the LLM testing procedure. The model prompt and design were aligned with the instructions for human participants, which started with explaining the dimension and listing the words to be rated. The LLMs would then provide ratings per word as required. **c**, The key domains studied span non-sensorimotor, sensory and motor domains, with specific example questions provided for each respective domain. The classification into 'non-sensorimotor' and 'sensorimotor' domains is based on whether the measures directly assess sensorimotor experiences (see above for more detailed information).

human participant research[1] for dimensions with strong model–human correlations. All *P* values reported below were corrected for multiple comparisons by controlling the false discovery rate (FDR)[43] (Methods).

Model–human alignment varies across domains

Dimension-wise correlations. To assess the similarity of model word ratings to human word ratings across each dimension, we calculated the Spearman rank correlation between model-generated and human-generated ratings at both the aggregate and individual levels. For the aggregated analyses, the model-generated ratings of each word were aggregated by averaging across the four rounds of each LLM, and human-generated ratings were averaged across individuals.

As shown in Fig. 2a,b, ChatGPT and Google LLMs exhibit strong correlations ($Rs > 0.50$) (see Supplementary Table 3 for additional statistics) with human ratings across most the non-sensorimotor dimensions. However, they show significantly weaker correlations in sensory and motor dimensions. This observation is supported by Mann–Whitney *U* tests comparing model–human similarities between the sensorimotor and the non-sensorimotor dimensions (GPT-4: $U(N_1 = 7, N_2 = 11) = 65.00$, $P = 0.018$, rank-biserial correlation (r_{rb}) = 0.69; GPT-3.5: $U(N_1 = 7, N_2 = 11) = 67.00$, $P = 0.010$, $r_{rb} = 0.74$; Gemini: $U(N_1 = 7, N_2 = 11) = 77.00$, $P < 0.001$, $r_{rb} = 0.10$; PaLM: $U(N_1 = 7, N_2 = 11) = 76.00$, $P < 0.001$, $r_{rb} = 0.97$).

We undertook an individual-level analysis to examine the similarity between human and model conceptual representations while considering individual variability. We constructed pairwise Spearman correlations for each pair of individual human participants (human–human) and each individual run of a model and each individual human participant (model–human). This process resulted in five distributions: human–human, GPT3.5–human, GPT4–human, PaLM–human and Gemini–human pairwise correlations. Using the human–human correlations as a benchmark for inter-person reliability, we asked: Are the responses from an individual model more or less similar to those of an individual human, as compared with the similarity between one human and another? Independent-sample *t*-tests assessed whether the distributions of model–human similarities significantly differ from human–human similarities, and Cohen's *d* was used to quantify the standardized difference between the two (see Supplementary Table 4 for additional statistics). A negative *d* value indicates that the model–human similarities are greater than the human–human similarities for that particular dimension and model.

Figure 3 presents model–human similarity distributions. The distributions marked with fill-in colours indicate comparisons where there is no evidence that a model's responses to a human are less similar than one human's responses to another human. The count of these marked distributions decreased from the non-sensorimotor domain

https://doi.org/10.1038/s41562-025-02203-8

Table 1 | Definitions of each dimension in Glasgow and Lancaster Norms

Norms	Domain	Dimension	Definition
Glasgow	Non-sensorimotor	Valence	Value or worth; representing something considered good or bad
		Dominance	The degree of control a word makes you feel
		Arousal	Excitement versus calmness
		Size	Dimensions, magnitude or extent of an object or concept that a word refers to
		Gender	How strongly its meaning is associated with male or female behaviour
		Concreteness	A measure of how concrete or abstract something is
		Imageability	How easy or difficult something is to imagine
Lancaster	Sensory	Haptic, auditory, olfactory, interoceptive, visual, gustatory	How much do you experience everyday concepts using six different perceptual senses
	Motor	Foot/leg, hand/arm, mouth/throat, torso, head excluding mouth	How much do you experience everyday concepts using actions from five different parts of the body

The two norms have incorporated different numbers of words/concepts in human judgments with approximately 4/5 of the Glasgow normed words overlapping with those in the Lancaster Norms (see Methods for details).

dimensions (16 marked model–human distributions out of 28 model–human distributions; 16/12) to the sensory (4/20) and further decreased in the motor dimensions (2/18), $\chi^2(2) = 15.49$, $P < 0.001$. Within the non-sensorimotor domain's seven dimensions, there is no credible evidence showing that GPT-4's model–human similarity distribution was significantly lower than the human–human distribution in all seven dimensions. For GPT-3.5, this held true in four dimensions, for Gemini in three dimensions and for PaLM in two dimensions. However, within the sensory domain's six dimensions, the count decreased to four for GPT-4 and to zero for all other models. Within the motor domain's five dimensions, the count further dropped to two for GPT-4 and remained at zero for the other models. These individual-level analyses reveal a growing divergence between models and humans from non-sensorimotor to sensorimotor dimensions.

RSA. While the above correlations capture model–human similarity over all words in each separate dimension, such dimension-wise analyses might overlook how different dimensions may jointly contribute to a word's overall conceptual representation and how different words are interconnected. For example, the concepts of 'pasta' and 'roses' might both receive high ratings for their olfactory qualities. However, 'pasta' is considered more similar to 'noodles' than to 'roses', not only because of its smell but also because of its visual appearance and taste. To address this issue, we adopt the RSA[44] to fully capture the complexities of word representations, where dimensions such as smell and visual appearance are considered jointly as part of a high-dimensional representation for each word.

RSA allows us to evaluate and compare how the geometric organization of concept words is aligned between models and humans across the non-sensorimotor, sensory and motor domains. To implement RSA (Fig. 4a), we represented each word as a vector separately within the non-sensorimotor, sensory and motor domains. The elements of these vectors were derived from the ratings of specific dimensions belonging to each respective domain. For example, the sensory vector for 'pasta' consists of ratings from six sensory dimensions (for example, haptic and auditory). We then constructed representational dissimilarity matrices (RDMs) by calculating the Euclidean distance between word vectors for each model and individual human, capturing word similarity relationships (for example, 'pasta' and 'noodles' are more similar than 'pasta' and 'roses'). The similarity between RDMs of each model and each individual human was calculated via the Spearman rank correlation. We thus obtained a distribution of similarities between all human participants and each model separately on each domain. We conducted two mixed-effects analyses of variance (ANOVAs) to statistically evaluate the model–human similarities across three domains, specifically to determine whether these similarities were lower in the sensory/motor domains compared with the non-sensorimotor domain. These analyses were performed separately for the ChatGPTs and Google LLMs, considering 'domain' and 'model' as two distinct factors.

For the ChatGPT models, a significant main effect of domain was observed ($F(2, 1{,}704) = 729.72$, $P < 0.001$, $\eta_p^2 = 0.46$). Both the sensory and motor domains showed significantly lower similarities compared with the non-sensorimotor domain. Specifically, the sensory domain had lower similarities than the non-sensorimotor domain ($t(1{,}526.5) = -2.93$, $P = 0.004$, $d = -0.13$, 95% confidence interval (CI) −0.03 to −0.01), and the motor domain was significantly lower than the non-sensorimotor domain ($t(1{,}731.8) = -44.49$, $P < 0.001$, $d = -1.87$, 95% CI −0.22 to −0.20). In addition, similarities in the motor domain were significantly lower than those in the sensory domain ($t(1{,}721.0) = -33.18$, $P < 0.001$, $d = -1.58$, 95% CI −0.21 to −0.19).

Google LLMs revealed similar results: a significant main effect of domain was observed ($F(2, 1{,}421) = 1{,}626.84$, $P < 0.001$, $\eta_p^2 = 0.70$). Both the sensory and motor domains showed significantly lower similarities compared with the non-sensorimotor domain. Specifically, the sensory domain had lower similarities than the non-sensorimotor domain ($t(892.4) = -49.22$, $P < 0.001$, $d = -2.46$, 95% CI −0.24 to −0.23), and the motor domain was significantly lower than the non-sensorimotor domain ($t(1{,}056.2) = -47.05$, $P < 0.001$, $d = -2.24$, 95% CI −0.23 to −0.21). Similarities in the motor domain were not significantly different from those in the sensory domain ($t(1{,}178.8) = 1.81$, $P = 0.077$, $d = 0.11$, 95% CI −0.00 to 0.02).

These results suggest that LLMs' conceptual representations and organizations of words align most closely with human representations in the non-sensorimotor domain, while alignments are weaker in the sensory domains and minimal in the motor domains. These observations are in line with earlier analyses, highlighting LLMs' progressively diminishing effectiveness in recovering human conceptual representations when they move towards more sensorimotor-related aspects of the representations (see Supplementary Table 5 for the descriptive statistics).

Linking additional visual training to model–human alignment

The increased disparity between model and human representations for more sensorimotor dimensions of conceptual representations suggests that grounding experience may be necessary to achieve human-like conceptual representation. Given this possibility, we pose a related question: What role do additional visual inputs play in conceptual formation within LLMs primarily trained on both language and visual inputs (for example, GPT-4 and Gemini, henceforth visual

https://doi.org/10.1038/s41562-025-02203-8

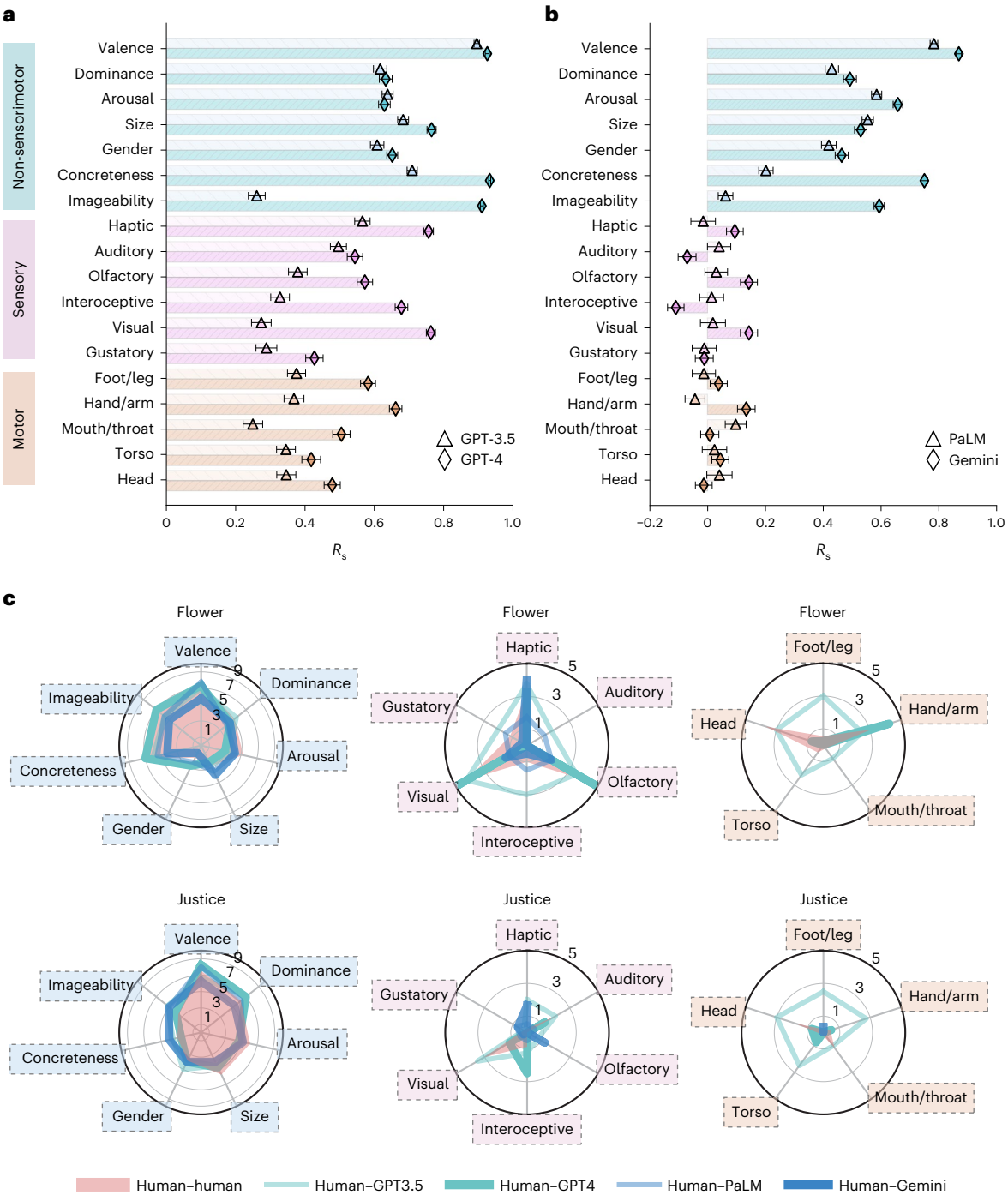

Fig. 2 | Aggregated results. a,**b**, Spearman correlations between the human-generated and LLM-generated ratings for all analysed words. The *x* axis represents the Spearman correlation coefficients between the aggregated word ratings generated by LLMs including GPT-3.5, GPT-4 (**a**), PaLM and Gemini (**b**) and the corresponding human ratings. The *y* axis lists the different dimensions being evaluated, along the non-sensorimotor, sensory and motor dimensions. The error bars depict the 95% confidence intervals, estimated by bootstrap resampling 1,000 samples of word ratings from aggregated human participants and LLMs. The central value represents the estimated correlation coefficient between the lower and upper confidence bounds. **c**, Radar plots showing the aggregated ratings of human, ChatGPT (GPT-3.5 and GPT-4) and Google LLMs (PaLM and Gemini) on each dimension for two individual concepts: 'flower' (a concrete word) and 'justice' (an abstract word). The numbers along the radial axis denote the rating ranges for these dimensions. Additional examples are provided in Supplementary Figs. 2 and 3.

LLMs) compared with those that have received input from only a single modality—language (for example, GPT-3.5 and PaLM, henceforth text-only LLMs)? In other words, is visual learning associated with the alignment of multimodal LLMs with human representations of sensorimotor concepts, and if so, how? Available information indicates that GPT-4 was pretrained with text and images[45], while Gemini was pretrained to integrate language data with a diverse array of visual inputs, including natural images, charts, screenshots, PDFs and videos[46]. By contrast, GPT-3.5 (ref. 47) and PaLM2 (ref. 48) were pretrained exclusively within the language domain.

Isolating the impact of visual training is challenging owing to limited access to the details of training in these state-of-the-art LLMs.

https://doi.org/10.1038/s41562-025-02203-8

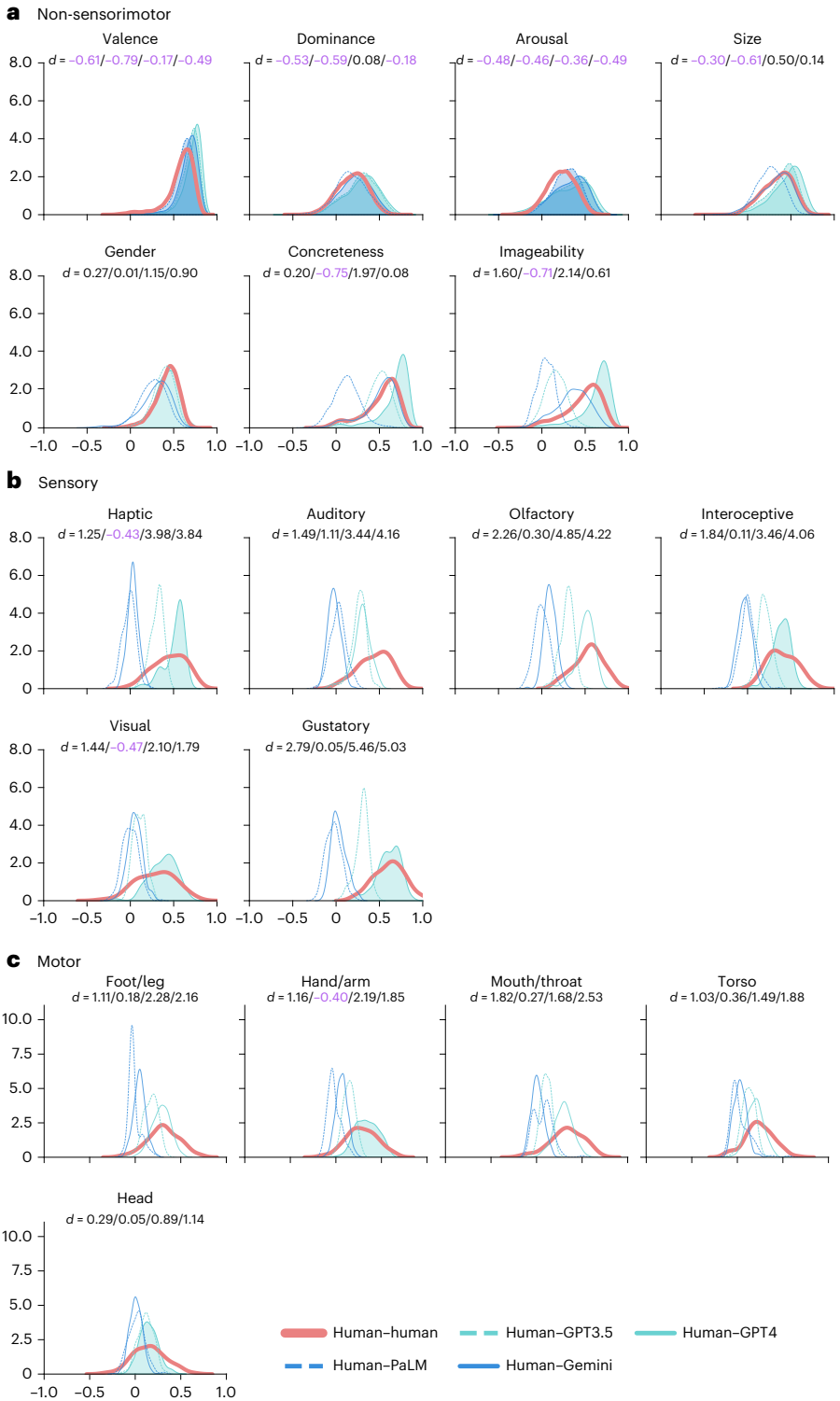

Fig. 3 | Individual analysis. a–c, The results for the individual-level pairwise correlation analysis for each dimension in the non-sensorimotor (**a**), sensory (**b**) and motor (**c**) domains. This analysis aims to examine the similarity between human and model conceptual representations while considering individual variability. The *x* axis represents the Spearman correlation coefficient, while the *y* axis shows the kernel density estimation of the correlation distributions. Notably, in the subplots for the motor dimensions, the *y* axis displays higher density peaks due to PaLM yielding model–human similarities clustered around zero. Cohen's *d* is reported for each dimension to quantify the standardized distance between the human–human and model–human correlation distributions. The *d* values for GPT-3.5, GPT-4, PaLM and Gemini models are presented between forward slashes ('/'), respectively. A negative *d* value, highlighted in purple, indicates that the model–human similarities are greater than the human–human similarities for that particular dimension and model. The distribution curves for human–human pairwise similarity, serving as benchmarks, are visually distinguished by the increased line thickness. When the colours are filled in model–human similarity distribution curves, they indicate that there is no credible evidence those model–human similarities are lower than human–human similarities (non-sensorimotor: 16 distributions out of 28 model–human distributions,16/12; sensory: 4/20; motor: 2/18). These filled-in curves highlight the dimensions and models where the model-generated ratings align closely with human ratings at the individual level. Here, the *P* values were assessed with two-sided *t*-tests and corrected for multiple comparisons using the FDR.

https://doi.org/10.1038/s41562-025-02203-8

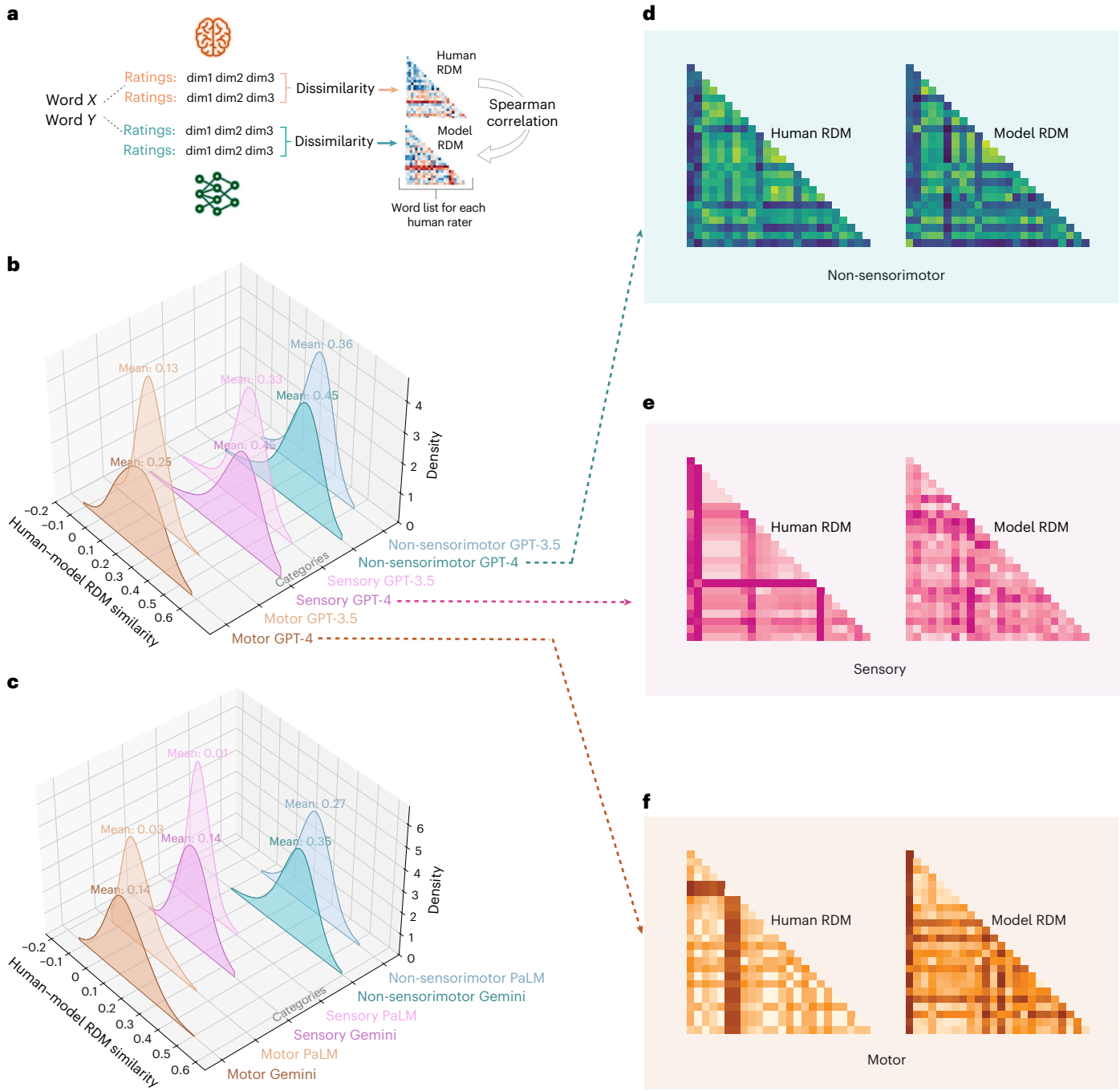

Fig. 4 | RSA. a, A schematic of the RSA: for each human rater and language model (GPT-3.5, GPT-4, PaLM and Gemini), the words were represented as separate vectors for the non-sensorimotor, sensory and motor domains. Icons from Flaticon.com[77]. The elements of these word vectors were derived from the ratings generated by humans or models for the dimensions belonging to each respective domain. The RDMs were then constructed by calculating the Euclidean distance between every pair of word rating vectors within each domain. Spearman correlations between these RDMs quantify the alignment of representational geometries, enabling comparison between human and model representations. **b**, Distributions of model–human RDM similarities for ChatGPT models (GPT-3.5 and GPT-4). The distributions of Spearman correlation coefficients for RDMs constructed upon individual human ratings and ChatGPT ratings for the same words across non-sensorimotor, sensory and motor domains are shown. The *x* axis represents Spearman correlation coefficients and *y* axis denotes the density of these coefficients. **c**, The distributions of model–human RDM similarities for Google LLMs (PaLM and Gemini). Similar to **b**, the distributions for human and Google LLM RDM alignment for the same words across the same three domains are displayed. Both **b** and **c** illustrate a trend that model–human RDM alignments decrease (with RDM similarities centralizing around smaller values) from non-sensorimotor to sensory and especially motor domains. **d**–**f**, Example RDMs: each RDM, constructed using 25 words, reflects pairwise similarities either based on human or GPT-4 ratings across non-sensorimotor (**d**), sensory (**e**) and motor (**f**) domains. The distinct patterns could be observed between the human RDM and GPT-4 RDM for the motor domain while for the non-sensorimotor domain, the human and GPT-4 model RDMs are much more similar.

Here, we piloted an analysis characterizing a potential association between the added visual learning and the difference in model–human alignment between visual and text-only LLMs. The rationale underlying this analysis is that if visual learning affects model alignment with human conceptual representations, this effect should be particularly noticeable in the visual dimension and in dimensions that involve some level of visual interpretation, such as imageability, which have been identified as having visual components in prior research (see refs. 49,50 for reviews). We quantified the visual association strength of each dimension by computing the Spearman correlation between each dimension and the visual dimension, using human rating data as reported in refs. 1,2. Higher absolute-value correlation coefficients indicate a stronger association with the visual dimension. For example, as illustrated in Fig. 5c, dimensions such as concreteness (R_s = 0.62, P < 0.001, 95% CI 0.60 to 0.64) and imageability (R_s = 0.69, P < 0.001, 95% CI 0.67 to 0.70) are strongly associated with the visual dimension, whereas dimensions such as gustatory (R_s = 0.01, P = 0.596, 95% CI −0.04 to 0.02) and torso (R = 0.02, P = 0.136, 95% CI −0.01 to 0.05) show minimal visual association.

To assess the change of visual LLMs over text-only LLMs in their alignment with human representations, we calculated the difference in model–human correlations (Fisher Z-transformed) between visual LLMs and text-only LLMs (GPT-4 versus GPT-3.5 for ChatGPTs and Gemini versus PaLM for Google LLMs) for each dimension. A higher value on a specific dimension indicates a stronger correlation between the visual LLM and human data compared with the text-only LLM for that dimension, as shown in Fig. 5a,b on the y axis. We then built separate linear regression models for ChatGPTs and Google LLMs, using the visual association strength as a predictor of the alignment change of the visual LLM over the text-only LLM. The results show that, for the ChatGPT models (Fig. 5a), the visual association strength was a positive predictor of the alignment change of GPT-4 over GPT-3.5 (B = 0.99, t(16) = 6.16, P < 0.001, 95% CI 0.65 to 1.33). Approximately 70% of the variance in the alignment change of GPT-4 over GPT-3.5 can be explained by the visual association strength of the dimensions (R^2 = 0.70), which suggests that visual inputs are a major factor in improving GPT-4's ability to align with human conceptual representations, highlighting the important role of visual learning in this context. For Google LLMs (Fig. 5b), the visual association strength also significantly positively predicted the alignment change of Gemini over PaLM (B = 0.43, t(16) = 2.38, P = 0.033, 95% CI 0.04 to 0.82). Approximately 26% of the variance in the alignment change of Gemini over PaLM can be explained by the visual association strength of the dimensions (R^2 = 0.26), similarly indicating the role of visual input in capturing human-like conceptual representations as seen in the ChatGPT models.

Validation of the results

Controlling for word concreteness. Will the divergence observed between LLMs and humans in sensory and motor domains persist when accounting for word concreteness? Prior work has suggested that LLMs may be capable of capturing the sensorimotor aspects of human representations of abstract words, which are purportedly less reliant on grounding[40,41]. To assess the potential influence of word concreteness on our findings, we employed two methods. First, we calculated the partial Spearman correlation between human and model ratings, controlling for word concreteness. Our analysis revealed a strong similarity (R_s = 0.93, P < 0.001, 95% CI 0.89 to 0.96) between these partial correlations and the original correlations reported earlier (Fig. 6a), suggesting that the pattern of our results remains even after adjusting for concreteness. Next, we implemented a bin analysis to explore variations in model–human correlations across different levels of word concreteness. We first sorted the words by concreteness values, then divided them into bins of 100 words each to explore variations in model–human correlations across different levels of word concreteness. For each bin, we recorded the median concreteness value and the corresponding model–human correlations for the words within the bin. Therefore, we obtained model–human correlations as a function of word concreteness for each model across various dimensions. The data were then analysed using a linear regression model, which considered concreteness values, model and domain as predictors of model–human correlations (Fisher Z-transformed). This model explained approximately 53% of the variance in model–human correlations. Our analysis did not identify credible evidence for a significant main effect of word concreteness on model–human correlations (Fig. 6b) (B = −0.00, t = −0.39, P = 0.785, 95% CI −0.03 to 0.02). Therefore, we found no credible evidence that word concreteness is associated with the alignment or divergence of models with human conceptual representations across non-sensorimotor, sensory and motor domains. That said, we did observe interaction effects between concreteness, domain and model, suggesting, for example, that model–human correlations may be stronger for more concrete words in the sensory domain (see Supplementary Information, section 5.4, for further analyses and Supplementary Information, section 7.2, for possible interpretations of these results).

Validating LLM responses. Because of the critical need for validity in LLM applications[18,42,51], we adhered to established human test validation methods[1,2]. We evaluated the ChatGPTs (GPT-3.5 and GPT-4) and Google LLMs (PaLM and Gemini) against a set of alternate norms that are related to the Glasgow and Lancaster measures.

For the Glasgow Norms, validation norms include dimensions of valence, arousal and dominance from ref. 34, imageability from ref. 52 and concreteness from ref. 14. For the Lancaster Norms, which lacks directly comparable validation norms, we included dimensions that are conceptually similar, such as taste and grasp[53]. In human ratings, taste is expected to strongly correlate with the gustatory dimension in Lancaster, while grasp shows moderate correlations with the hand/arm and haptic dimensions. We selected these validation norms for several reasons: (1) they are publicly accessible, (2) they have been widely used in human participant studies, lending to their validity and (3) they cover dimensions in either the Glasgow or Lancaster Norms where models show strong correlations (that is, R_s > 0.6) with human data.

As detailed in Table 2, we first evaluated models' responses on the validation norms, then computed Spearman correlations between humans and models for these norms. Subsequently, we calculated correlations for model ratings between the original Glasgow/Lancaster Norms and the validation norms. We observed that model–human correlations based on the validation norms—except for Gemini's performance on the arousal dimension—closely resembled those obtained from the Glasgow/Lancaster Norms. For instance, the correlation between human ratings and GPT-3.5 on valence was 0.83 (95% CI 0.82 to 0.84) in the validation norms, compared with 0.90 (95% CI 0.89 to 0.90) in the Glasgow Norms. Moreover, the correlation strength of ChatGPT ratings between the validation norms and the Glasgow/Lancaster norms is as high as the correlation strength of human ratings across these norm sets. For example, the correlation for GPT-4 ratings on the hand/arm dimension between the validation and the Lancaster norms was 0.68 (95% CI 0.62 to 0.73), compared with the 0.55 correlation of human ratings across these norms. Given the strong consistency observed across different models and dimensions, these results suggest that the main findings largely reflect the models' capabilities rather than reliance on specific prompts (see Supplementary Table 9 for additional statistics).

Discussion

In this study, we used LLMs to test the limits of conceptual knowledge acquisition by quantifying what aspects of human conceptual knowledge can or cannot be recovered solely from the language domain of learning or from a combination of language and visual domains. We found that learning constrained to the language domain captures

https://doi.org/10.1038/s41562-025-02203-8

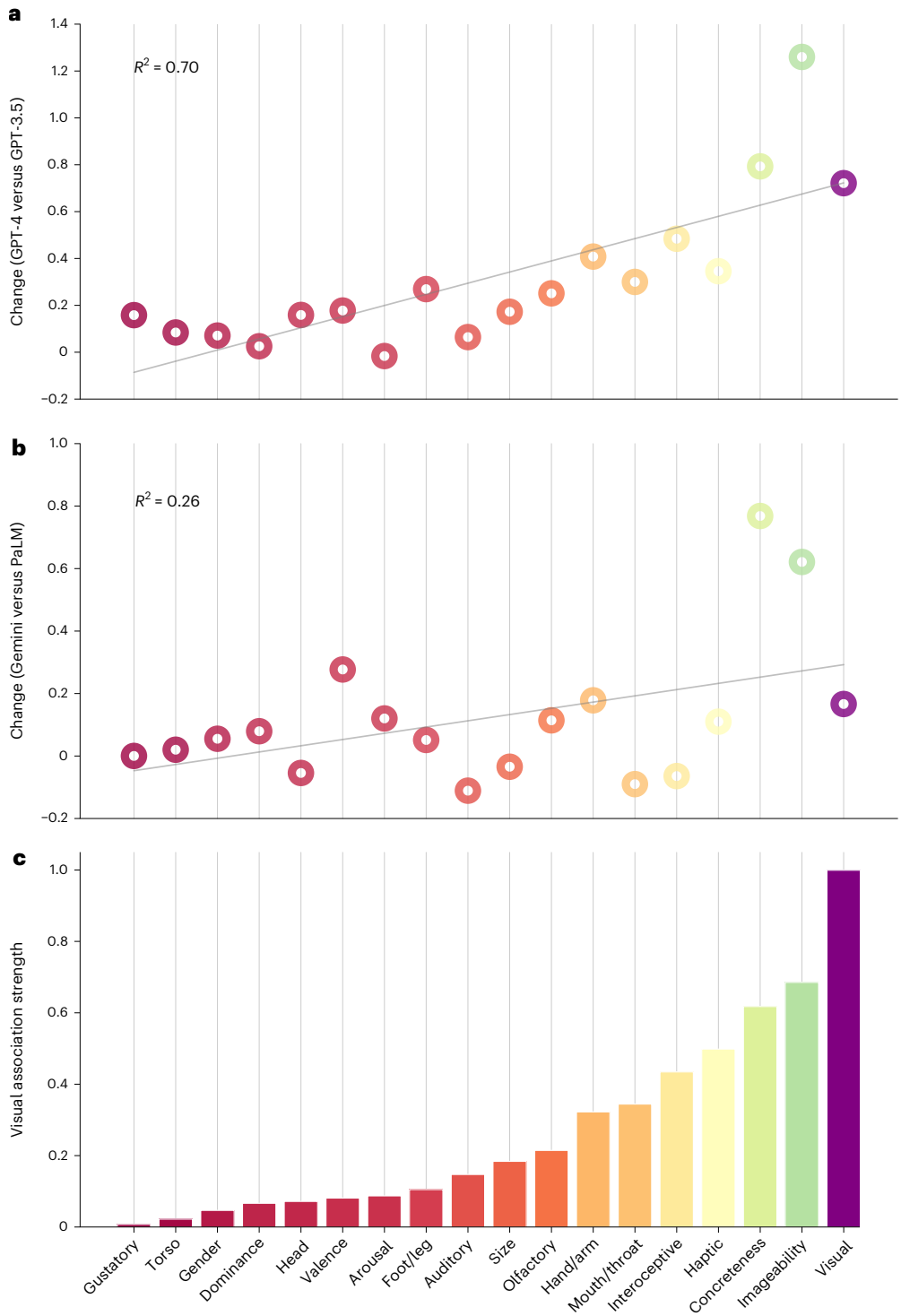

Fig. 5 | Visual domain analysis. a,**b**, Using the strength of visual correlation to predict the degree to which visual LLMs enhance alignment with human conceptual representations compared with their text-only counterparts. This comparison was made for ChatGPT models (GPT-4 versus GPT-3.5) (**a**) and Google LLMs (Gemini versus PaLM) (**b**). **c**, Visual association strength of each dimension: the absolute values of the Spearman correlation coefficients, based on human ratings[1,2], reflect the association strength of each dimension with the visual dimension. A higher coefficient signifies a stronger link to visual processing, such as the imageability and haptic dimensions. The *x* axis across all three subplots displays the dimensions, sorted by their visual association strength (as shown in **c**). In **a**, the *y* axis shows the difference in model–human correlations between GPT-4 and GPT-3.5. In **b**, similarly, the *y* axis denotes the difference in model–human correlations between Gemini and PaLM. In **c**, the *y* axis indicates the visual association strength of each dimension.

human-level conceptual representation in non-sensorimotor dimensions such as valence and emotional arousal but yields impoverished representation of sensorimotor knowledge. Our findings extend previous research on ungrounded artificial neural models[4–6] and congenitally blind and partially sighted people[7–10], which showed alignment with the conceptual representations of sighted human participants. By systematically examining conceptual representations across a spectrum from non-sensorimotor to sensorimotor domains and a wide

https://doi.org/10.1038/s41562-025-02203-8

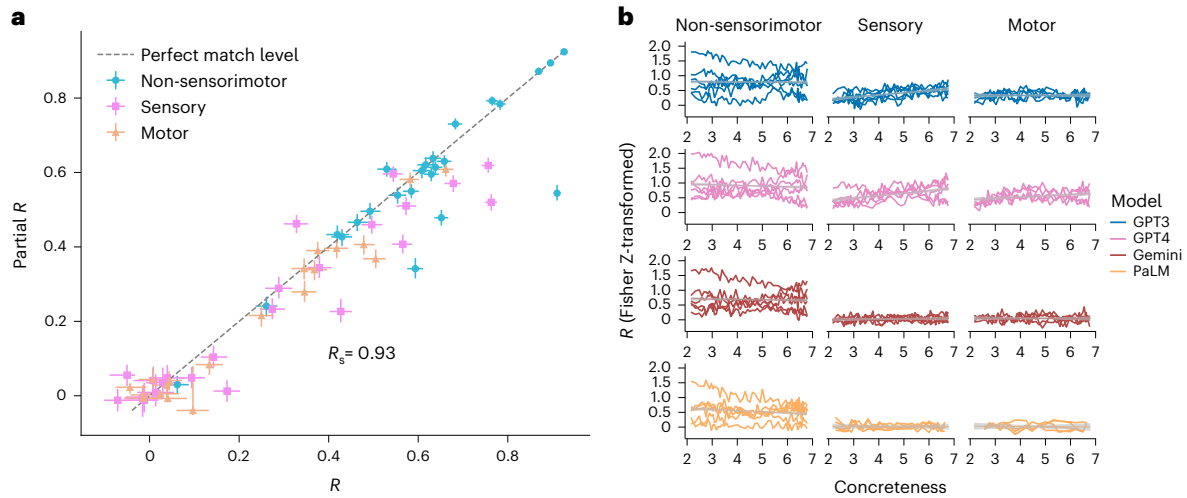

Fig. 6 | Concreteness analysis. a, The partial Spearman correlations between human and model ratings, controlling for word concreteness, are very similar (R_s = 0.93) to the original correlations. The dashed identity line indicates a perfect match, where the partial correlation value is exactly the same as the original correlation value. The error bars depict the 95% confidence intervals, estimated by bootstrap resampling 1,000 samples of ratings. The central value represents the estimated correlation coefficient between the lower and upper confidence bounds. **b**, A bin analysis of the correlations between human and model ratings across different levels of word concreteness for each model (GPT-3.5, GPT-4, PaLM and Gemini) and domain (non-sensorimotor, sensory and motor). The words were first sorted by concreteness values and then divided into bins of 100 words each. The *x* axis shows the median concreteness value within each bin. The *y* axis denotes model–human Spearman correlations (Fisher *Z*-transformed) for words within the bin. The fit line for each panel represents the prediction from the regression model for each domain and each model.

Table 2 | Spearman correlations between LLM–human ratings for the original and validation norms, and between original–validation norms for LLM and human ratings

Model	Dimension	LLM–human		Original–validation	
		Original	Validation	LLM	Human
GPT-3.5	Valence	0.90	0.83	0.90	0.93
	Dominance	0.62	0.66	0.82	0.69
	Arousal	0.64	0.47	0.55	0.62
	Concreteness	0.71	0.63	0.61	0.93
GPT-4	Valence	0.93	0.88	0.92	0.93
	Dominance	0.63	0.67	0.86	0.69
	Arousal	0.64	0.43	0.54	0.62
	Concreteness	0.93	0.87	0.88	0.93
	Imageability	0.91	0.77	0.83	0.89
	Haptic	0.76	0.88	0.55	0.55
	Hand/arm	0.66	0.88	0.68	0.55
PaLM	valence	0.78	0.44	0.42	0.91
Gemini	Valence	0.87	0.83	0.82	0.91
	Arousal	0.66	0.15	0.39	0.60
	Concreteness	0.75	0.66	0.58	0.93

Original norms denote Glasgow or Lancaster. All correlations were significant (P<0.001).

range of concepts, we found a gradual decrease in similarity between LLM-derived and human-derived representations, with stronger disparity in sensorimotor domains. These results offer insights into the extent to which language can shape complex concepts and underscore the importance of multimodal inputs for LLMs emulating human-level conceptual knowledge.

In light of the ongoing debate about the necessity of embodied grounding for achieving human-level conceptual representation[26,32], The present study suggests that while some aspects of conceptual representations may be detached from sensorimotory experience, a considerable degree of sensorimotor input appears essential. Take the concept of 'flower', for instance. Language can capture certain conceptual connotations of 'flower' insofar as they emerge from distributional relationships among words in context (for example, positive emotional valence may arise from 'this flower smells joyous'). However, the sensorimotor experience of 'flower' may cut across linguistic contexts and may implicitly shape our conceptual knowledge, to form diverse relationships across objects and experiences in the world around us. From the intense aroma of a flower, the vivid silky touch when we caress petals, to the profound visual aesthetic sensation, human representation of 'flower' binds these diverse experiences and interactions into a coherent category. This type of associative perceptual learning, where a concept becomes a nexus of interconnected meanings and sensation strengths, may be difficult to achieve through language alone. Real-world interactions, similar to those in human experiences, are probably essential for comprehensive sensory perception, physical action and perceptual representation of concepts[54].

Intriguingly, we found greater discrepancies between human and LLM ratings for motor-related dimensions than for sensory dimensions—an area underexplored in prior studies. Two explanations for this finding are: (1) motor aspects are less frequently described in language, making them harder for LLMs to learn from language, as noted by ref. 55, and (2) motor representations rely more on embodied experiences, unlike sensory concepts such as colour, which can be learned through language[9]. Motor cortex lesions also impair action-word processing, underscoring the need for embodiment[56]. This further highlights LLMs' limitations in representing motor concepts owing to the lack of physical commonsense and action-related input.

The current study exemplifies the potential benefits of multimodal learning where 'the whole is greater than the sum of its parts', showing how the integration of multimodal inputs can potentially lead to a more human-like representation than what each modality could offer independently. We found that LLMs incorporating visual inputs align better with human representations in visual as well as visual-related dimensions, such as haptics and imageability. This representational transfer is well observed in humans[57,58]. For instance, humans can acquire

object-shape knowledge through both visual and tactile experiences[57], and brain activation in the lateral occipital complex was observed during both seeing and touching objects[59]. Akin to humans, given the architecture and learning mechanisms of visual LLMs, where representations are encoded in a continuous, high-dimensional embedding space, inputs from multiple modalities may fuse or shift embeddings in this space. The smooth, continuous structure of this embedding space may underlie our observation that knowledge derived from one modality seems to spread across other related modalities[60–62]. Further along this vein, our study points to the possibility that models may be able to approximate human-like conceptual representations even without full sensorimotor experience; partial access could suffice to span much of human experience[15]. This insight may also shed light on why similar representations were observed between congenitally blind and partially sighted people and normally sighted people[8–10]. Future research should explore the extent of sensory access needed in multimodal models and the limits of knowledge transfer across different domains. The continued development of LLMs towards integrating additional modalities—as seen in multimodal speech and text processing in Whisper[63] and embodied vision-language-action models such as RT-2 (ref. 64)—opens exciting prospects for further understanding and harnessing the potential of multimodal learning. We envision a future where LLMs are augmented with sensor data and robotics to actively make inferences about and act upon the physical world[16,17]. These advances may catalyse LLMs to truly embrace embodied artificial representation that mirrors the complexity and richness of human cognition[17,29]. Within this perspective, our findings may contribute to the trajectory of training data improvement and multimodal integration.

To what extent do LLMs inform us about human cognition? LLMs revive the old debate about whether and how distributional relationships in language-domain learning can scaffold a wide range of semantic processes, reflecting the richness of linguistic inputs in shaping human knowledge[20,65,66]. At the same time, their limitation in capturing human-like sensorimotor conceptual understanding via textual data and incomplete sensorimotor input also delineates the boundary of language-domain training and underscores the importance of grounding for human conceptual knowledge[15]. In this light, LLMs offer a valuable 'how-possibly' model of human cognition. Nonetheless, we acknowledge that most work on the parallels between LLMs and human language processing (for example, refs. 21,22), including the current work, has been confined to the English language. This constitutes a limitation of our study, as language structure, embodiment effects and neural processing could differ across languages. However, findings like those of ref. 67 suggest that motor verbs in French and German elicited similar motor-related brain activations compared with non-motor verbs, indicating that our English-based findings might generalize to other languages. Future studies should explore using diverse languages to validate and expand these insights (see Supplementary Information, section 7.4, for a further discussion on the cognitive plausibility of LLMs).

It is worth noting that LLMs involve diverse learning techniques, which adds complexity to their learning dynamics and makes it valuable to explore how each technique contributes to the final outcomes. For example, GPT-3.5 is pretrained using next-token prediction within text sequences and then further refined through two methods: supervised learning, where human-labelled data specifies the correct output, and reinforcement learning with human feedback (RLHF), which enables the model to improve by interacting with the external environment indirectly[47]. These two techniques could also bring in non-linguistic knowledge, as humans provide labels or feedback based on some non-linguistic resources. We believe these techniques do not alter our main findings, as RLHF is constrained by mechanisms such as Kullback–Leibler (KL) divergence penalties[68], which prevent the fine-tuned model from diverging substantially from the pretrained version. Although RLHF may indirectly introduce human preferences based on real-world non-linguistic experiences, the model's learning remains primarily driven by linguistic input. However, a key limitation of our study is the proprietary nature of these large models, which makes it challenging to conduct open scientific research and fully understand the individual effects of each learning approach, including the specific impacts of RLHF. This opacity hinders our ability to dissect how different techniques influence the model's behaviour. Therefore, our conclusions about how LLMs acquire and process language and embodied experiences should be interpreted with these considerations in mind, underscoring the need for greater transparency in LLM research to enable more systematic investigations. Notably, DeepSeek-V3 (ref. 69), a recent high-performance open-source model with various post-training optimization, shows performance comparable to GPT-4 (Supplementary Fig. 5), further supporting our conclusion: LLMs capture non-sensorimotor semantics well but still struggle with nuanced sensory and motor features of words and concepts. Future research could focus on smaller, more accessible models to test and compare the roles of various learning techniques, such as prediction-based learning, supervised learning and interaction-based reinforcement learning (see ref. 54 for an example). This is especially important given that leveraging multiple knowledge resources and interacting with the environment have long been recognized as crucial and efficient mechanisms in human language and concept development[16,17,70].

We note that, although LLMs can approximate certain aspects of conceptual representation, particularly in non-sensorimotor and occasionally bodily dimensions, they obtain this by consuming vast amounts of text—orders of magnitude larger than the volume of language a human is exposed to in their entire lifetime—and operating with extremely high complexity driven by billions of parameter settings[20]. This suggests that, while in the limit multimodal knowledge can be synthesized from language alone, this kind of learning is inefficient. By contrast, human learning and knowledge representation are both inherently multimodal and embodied, and interactive from the outset[15]. After all, when thinking of flowers, what comes to your mind is not merely their names but the vivid symphony in which sight, touch, scent and all your past sensorimotor experiences intertwine with profound emotions evoked—an experience far richer than words alone can hold.

Methods

Inclusion and ethics

The study involves the collection of data from LLMs and the use of secondary human-participant data[1,2]. For the human-participant data, ref. 1 noted that the study followed the ethical guidelines and protocols established by the British Psychological Society. Ethical approval for the study reported in ref. 2 was granted by the Lancaster University Research Ethics Committee.

Psycholinguistic norms

We used the Glasgow Norms[1] and the Lancaster Sensorimotor Norms (henceforth the Lancaster Norms[2]) as human psycholinguistic word rating norms (see Table 1 for their dimensions). Together, the two norms offer comprehensive coverage of the included dimensions, both of which cover a large number of words. The Glasgow Norms collected data from 829 human participants, including 599 female and 230 male participants in terms of gender. The original publication did not specify whether sex and/or gender was determined by self-report or assignment. Participants ranged in age from 16 to 73 years, with a mean of 21.7 years (standard deviation (s.d.) of 7.4). The average age was 21.5 years (s.d. of 7.6) for female participants and 22.3 years (s.d. of 6.9) for male participants. The Lancaster Norms collected data from 3,500 human participants, including 1,644 female and 1,823 male participants. A total of 12 participants chose not to disclose their gender, and the gender information was missing for 21 participants. The average age of all participants was 34.9 years (s.d. of 10.3).

The Glasgow Norms consist of normative ratings for 5,553 English words across nine dimensions, collected from native English speakers within the University of Glasgow community, UK[1]. We selected the Glasgow Norms owing to its large-scale data and highly standardized data collection process: the same participants rated all dimensions for any given subset of words, with an average of 33 participants per word. The nine dimensions include emotional arousal, valence, dominance, concreteness, imageability, size, gender association, familiarity and age of acquisition. In our study, we excluded familiarity and age of acquisition, as familiarity is less dependent on semantic and conceptual representation[71] and, therefore, less relevant to our research focus, while age of acquisition is neither central to our focus and nor a valid question for LLMs to answer. The validity of the Glasgow Norms has been demonstrated through strong correlations with 18 different sets of other psycholinguistic norms. Scott et al.[1] conducted principal component analyses and identified three main categories underlying these dimensions: emotion (valence and dominance), salience (arousal, size and gender) and mental visualization (concreteness and imageability). We adopt their validated structure for categorizing those dimensions.

The Lancaster Norms present multidimensional measures encompassing sensory and motor strengths for approximately 40,000 English words, collected from experienced users on Amazon's Mechanical Turk platform[2]. These norms include six sensory dimensions (haptic, auditory, olfactory, interoceptive, visual and gustatory) and five motor dimensions (foot/leg, hand/arm, mouth/throat, torso and head excluding mouth). The sensorimotor properties of words are considered highly embodied, as they require human raters to utilize their everyday perceptual senses and bodily experiences to gauge each word. The data were collected from 3,500 unique participants, with each participant rating on average 7.12 lists for either the sensory or motor dimensions. Each list comprised 58 words, including 48 target words, 5 control words and 5 calibration words. The fixed sets of five control words were randomly interspersed to each item list to ensure the quality of participants' ratings, and the five calibration words were presented at the beginning of each item list to introduce participants to unambiguous examples for rating. The Lancaster Norms were chosen primarily because they provide a detailed and comprehensive representation of a word's perceived sensorimotor strengths across 11 dimensions, covering all senses and the five most common action effectors. The norms exhibit high reliability, displaying substantial consistency across all dimensions, and their validity is demonstrated by their ability to accurately represent lexical decision-making behaviour from two distinct databases[2].

There are differences between the Glasgow and Lancaster Norms in rater demographics. To ensure that any observed differences between the sensorimotor and non-sensorimotor dimensions are attributable to the intended dimensions rather than these demographic differences, it is essential to confirm the validity of both the norms and our model's responses to them. The validity of the Glasgow and Lancaster Norms has been well established[1,2], and the validity of model responses to them is reported in the 'Validation of results' section in Results.

We adhered to the design of the human-participant data collection (Fig. 1b)[1,2]. For the Glasgow measures, the 5,553 words were divided into 40 lists, with 8 lists containing 101 words per list and 32 lists containing 150 words per list. The models rated all words in a list for one dimension before moving on to the next dimension and so forth. The order of words within each dimension and the order of dimensions within each testing round was randomized. For the Lancaster measures, there are in total 39,707 available words with cleaned and validated sensorimotor ratings. We first extracted 4,442 words overlapping with the 5,553 words in the Glasgow measures. Following the practice in the Lancaster Norms, we obtained the frequency and concreteness measures[14] of these 4,442 words and attempted to perform quantile splits over them to generate item lists that maximally resemble those in the Lancaster Norms. However, since more than 95% of the 4,442 words have a 'percentage of being known' greater than 95%, we considered the majority of these words to be recognizable by human raters. Thus, we did not perform a quantile split of these words over word frequency. We instead implemented a quantile split based on their concreteness ratings with four quantile bins in the intervals 1.19–2.46, 2.46–3.61, 3.61–4.57 and 4.57–5.00.

Next, we generated four sublists based on the concreteness rating quantile split and randomly selected 12 words from each sublist without replacement to create 48 words for each item list. We further appended the five calibration words (sensory dimensions: account, breath, echo, hungry and liquid; motor dimensions: shell, tourism, driving, breathe and listen) to the beginning of each list. Finally, we randomly inserted five control words (sensory dimensions: grass, honey, laughing, noisy and republic; motor dimensions: bite, enduring, moving, stare and vintage) into these lists to form 93 complete item lists, each containing 58 words ready to be rated separately for sensory and motor dimensions. The order of words within each item list and the order of dimensions to rate for each round were randomized.

Models

We employed the gpt-3.5-turbo-0301 and gpt-4 (collected between 28 May and 11 June 2023) from the OpenAI API for GPT-3.5 and GPT-4 and the PaLM2 (PaLM2 ratings of 2,474 words for the sensory dimensions were collected and PaLM2 ratings of 4,095 words for motor dimensions were collected since PaLM2 failed at returning ratings for several lists of words in each model run) and gemini-1.0-pro from the Google API for PaLM and Gemini. The selection of parameters in our study was based on methodological considerations aimed at optimizing the accuracy and consistency of the model outputs. The temperature parameter was set to 0, following recommendations described previously[21,22]) to ensure deterministic, consistent responses without random variations. The maximum token length was set to the upper limits permitted—2,048 tokens for GPT-3.5, GPT-4 and Gemini and 1,024 tokens for PaLM—to avoid truncating responses. To enhance the reliability of our results, we implemented four rounds of testing for each model. This approach allowed us to cross-verify the consistency of the outputs across multiple iterations (see Supplementary Information, section 1, for the agreement between these rounds).

Testing procedure

The model prompt to ChatGPT was kept identical to the instructions that human participants received. However, we made minor adjustments to the prompt to ensure that the responses followed the expected format (for example, word – rating). When given testing items from the Lancaster Norms, the model consistently responded that it does not possess a biological body and, therefore, cannot experience the word through sensing or moving. To address this, we modified the instruction from 'to what extent do you experience' to 'to what extent do human beings experience', and we applied the same changes to the Glasgow Norms for consistency. Although the LLM is asked to respond on the basis of human experience, it is still utilizing its internal representations to provide answers. These representations are derived from extensive training on human-generated text, which makes the responses valid as a reflection of the collective conceptual representation of humans.

The images or tables used in human-participants tasks were converted to text format. Moreover, the online rating portal of the Lancaster Norms used a graphic demonstration of the five body parts for the action-executing effector ratings. Because GPT-3.5 and PaLM do not support such visual inputs in the prompts, we decided instead to describe these five body parts with words in the prompts for all of the models (see Supplementary Information, section 2, for a comparison between the instructions given to human participants and the adapted version provided to the models).

Words analysed

For more uniform comparisons across various dimensions, we restricted our analysis of sensory and motor domains to words common to both the Glasgow and Lancaster Norms (4,442 words). Still, we retained the full Glasgow Norms (5,553 words) for the non-sensorimotor domain. Each of the overlapped 4,442 words has corresponding ratings across all evaluated dimensions. In occasional instances, models classified certain words as 'unknown' or unable to gauge (PaLM failed at generating all 4,442 words as detailed above). These words are typically those that may contravene content policies or that models such as PaLM struggle to interpret. In line with the practices described in refs. 1,2, such data points (that is, scores from individual runs) were excluded from the data analyses. To ensure that our results were not biased by words present in the Glasgow Norms but are not included in the Lancaster Norms, we conducted separate tests using only the fully overlapping concepts (4/5 of the Glasgow Norms) and found highly consistent results (Supplementary Information, section 6).

Individual-level pairwise correlations

For individual-level analysis, we computed pairwise Spearman correlations for each pair of individual human participants and between each human and individual runs of GPT-3.5, GPT-4, Gemini and PaLM. In the human–human correlations, each participant evaluated only a subset of words. In the Glasgow Norms, participants rated one of either 8 lists (comprising 808 words in total, with 101 words per list) or 32 lists (from a pool of 4,800 words, with 150 words per list). Each list received ratings from 32–36 participants, and there was no overlap in words across different lists. The pairwise correlations were calculated within each list, and these were aggregated, resulting in a total of 22,730 pairs for constructing the overall distribution for each dimension in the Glasgow Norms.

In the Lancaster Norms, the sensory component involved 2,625 participants (averaging 5.99 lists each) and the motor component had 1,933 participants (averaging 8.67 lists each). Each list included 48 test items, along with a constant set of five calibration and five control words, totalling 58 items per list. Given the larger pool of 40,000 words in the Lancaster Norms, the subset of 4,442 words resulted in some participants rating few items. To maintain a sufficient sample size for correlation calculations, we iterated through pairs of participants and included those with ratings for over 50 common words. This approach yielded 105 pairs for every sensory dimension and 196 pairs for every motor dimension, from which we constructed the correlation distributions.

In the human model correlations, we generated pairs by matching each model run (out of four total runs) with individual human participants across different lists. This approach yielded 5,476 pairs for the Glasgow Norms. For the Lancaster Norms, we paired humans and models based on having ratings for over 50 common words, mirroring the approach used in constructing human–human pairs. This process resulted in a total of 224 pairs for each sensory dimension and 440 pairs for each motor dimension, forming the basis for the correlation distributions.

RSA

For the RSA analysis, we first iterated through human rating data from the Glasgow and Lancaster Norms, extracting ratings across the non-sensorimotor, sensory and motor domains for lists of words rated by individual human participants. Each word was represented by a vector containing human ratings for each domain (for example, a vector for the sensory domain included ratings from six typical senses). Next, with these vectors, we built RDMs by calculating pairwise Euclidean distances between words for each list rated by individual human participants. This process was repeated for all three domains. Finally, we compared the RDMs derived from individual human ratings with model RDMs. The model RDMs were constructed using averaged ratings across four runs generated by the GPT models and Google models for the same words in each human word list. The comparison between human and model RDMs was conducted using Spearman correlation.

To ensure consistency and maintain a sufficient sample size for the RSA analysis, we only paired human and model data that had at least 50 shared words in each of the non-sensorimotor, sensory and motor domains for each model. As a result, we retained 829 pairs of RDMs from the Glasgow Norms for the non-sensorimotor domain RSA, applicable to both GPT and Google models. For the Lancaster Norms, we retained 435 pairs of RDMs for the sensory domain RSA and 443 pairs for the motor domain RSA with the GPT models. For the Google models, we retained 272 pairs of RDMs for the sensory domain RSA and 323 pairs for the motor domain RSA.

Measuring similarities

Spearman correlations were used for most similarity measurements, a common practice in many previous studies[44,72], as they are known to be robust with respect to outliers. For better presentation of correlations, we follow standard benchmarks: values under 0.10 are negligible, 0.10 is small, 0.30 medium and 0.50 or higher large[73]. We used Euclidean distance instead of Spearman correlations in only one case: when constructing RDMs for each word pair separately for humans and models. This decision was due to the small dimensionality of word rating vectors in each domain—$N = 7$ for the non-sensorimotor domain, $N = 6$ for the sensory domain and $N = 5$ for the motor domain. In such small dimensions, the ranking process used in Spearman correlations becomes less reliable and more sensitive to small variations in data[74], diminishing the ability to distinguish between different levels of similarity as the number of elements (that is, ranks) decreases. Conversely, Euclidean distance measures the 'straight line' distance between points in multidimensional space and is based on actual values rather than rankings. This measure tends to be more stable in cases of small dimensionality because it does not rely on rank-order relationships but on the actual differences in values across dimensions. However, because Euclidean distance is sensitive to the scale of the data, we normalized the rating values of each dimension using a z-score before obtaining vectors for each word in each domain.

Significance testing

All reported statistics are based on two-sided tests. The P values reported in each section were corrected for multiple comparisons using the FDR method[43]. In the Results section, the correction accounted for 157 tests, encompassing 72 aggregated-level dimension-wise correlations, 72 individual-level t-test comparisons between human–human distributions and model–human distributions, 4 Mann–Whitney U tests, 1 χ^2 test and 8 comparisons for RSA analyses. For the 'Linking additional visual training to model–human alignment' section in Results, the correction covered 20 tests, including 18 correlations between each dimension and the visual dimension and 2 t-tests assessing the predictive effect of visual correlation strength. In the 'Validation of results' section in Results, the correction was applied over 79 tests, which include 1 test for the comparison between the partial and the original correlation values, 47 tests for the bin analysis and 30 tests for using validation norms to validate model responses.

For the dimension-wise correlation analyses in the Results section on aggregated model and human ratings, we utilized the Mann–Whitney U test for independent-sample non-parametric comparisons of model–human similarities between sensorimotor and non-sensorimotor domains. This approach was selected owing to the small sample size (in this context, the number of dimensions within each domain) and the violation of normal distribution assumptions by the data. Non-parametric tests, although generally less powerful than parametric tests, are robust to outliers in such scenarios. Moreover, we reported the effect size using the rank-biserial correlation (r_{rb}),

a common measure for non-parametric tests. For the dimension-wise correlation analyses in the individual-level analysis, we utilized independent-sample *t*-tests to determine whether the distributions of model–human similarities differed significantly from human–human similarities. This approach was based on the assumption of normality and the presence of a large sample size. We conducted a separate *t*-test for each dimension. To organize the results of multiple *t*-tests effectively, we counted, within each domain, the number of instances where the distributions of model–human similarities were not significantly lower than those of human–human similarities, as indicated by the *t*-tests. A χ^2 test of independence was then performed to assess whether the counts varied significantly across the domains (non-sensorimotor, sensory or motor).

For the RSA analysis ('RSA' section in Results), after obtaining distributions of similarities between all human subjects and each model separately for each domain, we conducted a 3 × 2 (domain levels by models, respectively) two-way ANOVA for each set of models—ChatGPTs and Google LLMs—separately. This separation was to assess the consistency of main effects of domain across the two LLM families. Owing to violations of the equal variances assumption, we used the Satterthwaite's method for the ANOVA tests and applied Welch's corrections for post hoc pairwise comparisons. In the linear regression analyses ('Linking additional visual training to model–human alignment' section and 'Validation of results' section in Results), we conducted analyses after checking the assumptions of linearity, independence of residuals and normality. While the regression models in the 'Linking additional visual training to model–human alignment' section in Results meet all assumptions, the model in the 'Validation of results' section in Results shows a slight violation of the normality of residuals assumption, as indicated by the Normal *Q*–*Q* plot. To ensure the reliability of the results in the 'Validation of results' section in Results, we conducted an additional Bayesian linear regression analysis for cross-validation (Supplementary Information, section 5.4). In the 'Linking additional visual training to model–human alignment' section in Results, the Fisher *Z*-transformation was applied to the Spearman *R* values to measure the difference between two correlation coefficients, a practice that is justified by ref. 75.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data obtained from ChatGPTs and Google LLMs are publicly available at https://osf.io/kguwd/. The human dataset of the Glasgow Norms is from ref. 1, with word-level data accessible at https://doi.org/10.3758/s13428-018-1099-3 (ref. 1). The corresponding trial-level data was kindly provided by Sara Sereno and Jack Taylor. The Lancaster Norms is from ref. 2, and the data including both word-level and trial-level can be found at https://embodiedcognitionlab.shinyapps.io/sensorimotor_norms/ (ref. 76). The validation norms datasets are openly available via the following links: the datasets of valence, arousal and dominance at https://link.springer.com/article/10.3758/s13428-012-0314-x#SecESM1 (ref. 34), the imageability norms at https://link.springer.com/article/10.3758/BF03195585#SecESM1 (ref. 52), the concreteness norms at https://link.springer.com/article/10.3758/s13428-013-0403-5#MOESM1 (ref. 4) and the perceptual strength norms at https://link.springer.com/article/10.3758/s13428-012-0215-z#SecESM1 (ref. 3). Source data are provided with this paper.

Code availability

The data collection and analyses were conducted using Python and R. All code is publicly available at https://osf.io/kguwd/. In addition, we developed an analysis pipeline that enables researchers to examine their models of interest. As new models continue to emerge, we will regularly update the repository to ensure its ongoing relevance for the research community. The pipeline and associated resources are also accessible via GitHub at https://github.com/qxu1994/LLM_grounding.

References

1. Scott, G. G., Keitel, A., Becirspahic, M., Yao, B. & Sereno, S. C. The Glasgow norms: ratings of 5,500 words on nine scales. *Behav. Res. Methods* **51**, 1258–1270 (2019).
2. Lynott, D., Connell, L., Brysbaert, M., Brand, J. & Carney, J. The Lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods* **52**, 1271–1291 (2020).
3. Barsalou, L. W. Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
4. Patel, R. & Pavlick, E. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations* (2022).
5. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat. Hum. Behav.* **6**, 975–987 (2022).
6. Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N. & Griffiths, T. L. Large language models predict human sensory judgments across six modalities. *Sci. Rep.* **14**, 21445 (2024).
7. Bi, Y. Dual coding of knowledge in the human brain. *Trends Cogn. Sci.* **25**, 883–895 (2021).
8. Wang, X., Men, W., Gao, J., Caramazza, A. & Bi, Y. Two forms of knowledge representations in the human brain. *Neuron* **107**, 383–393 (2020).
9. Kim, J. S., Aheimer, B., Montané Manrara, V. & Bedny, M. Shared understanding of color among sighted and blind adults. *Proc. Natl Acad. Sci. USA* **118**, e2020192118 (2021).
10. Bottini, R. et al. Brain regions involved in conceptual retrieval in sighted and blind people. *J. Cogn. Neurosci.* **32**, 1009–1025 (2020).
11. Banks, B. & Connell, L. Multi-dimensional sensorimotor grounding of concrete and abstract categories. *Philos. Trans. R. Soc. B* **378**, 20210366 (2023).
12. Pexman, P. M., Diveica, V. & Binney, R. J. Social semantics: the organization and grounding of abstract concepts. *Philos. Trans. R. Soc. B* **378**, 20210363 (2023).
13. Lenci, A., Baroni, M., Cazzolli, G. & Marotta, G. Blind: a set of semantic feature norms from the congenitally blind. *Behav. Res. Methods* **45**, 1218–1233 (2013).
14. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
15. Dove, G. Symbol ungrounding: what the successes (and failures) of large language models reveal about human cognition. *Philos. Trans. B* **379**, 20230149 (2024).
16. Pezzulo, G., Parr, T., Cisek, P., Clark, A. & Friston, K. Generating meaning: active inference and the scope and limits of passive ai. *Trends Cogn. Sci.* **28**, 97–112 (2024).
17. Borghi, A. M., De Livio, C., Mannella, F., Tummolini, L. & Nolfi, S. Exploring the prospects and challenges of large language models for language learning and production. *Sist. Intell.* **35**, 361–378 (2023).
18. Frank, M. C. Large language models as models of human cognition. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/wxt69 (2023).
19. Blank, I. A. What are large language models supposed to model? *Trends Cog. Sci.* **27**, 987–989 (2023).
20. Connell, L. & Lynott, D. What can language models tell us about human cognition? *Curr. Dir. Psychol. Sci.* **33**, 181–189 (2024).
21. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).

22. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl Acad. Sci. USA* **121**, e2405460121 (2024).
23. Cai, Z. G., Haslett, D. A., Duan, X., Wang, S. & Pickering, M. J. Do large language models resemble humans in language use? *In Proc. Workshop on Cognitive Modeling and Computational Linguistics* 37–56 (ACL, 2024).
24. Piantadosi, S. in *From Fieldwork to Linguistic Theory* (eds Gibson, E. & Poliak, M.) 353–414 (Language Science, 2024).
25. Piantadosi, S. T. & Hill, F. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)* (NeurIPS, 2022).
26. Jacob, B. et al. Do large language models need sensory grounding for meaning and understanding? *The Philosophy of Deep Learning* https://phildeeplearning.github.io/ (2023).
27. Mitchell, M. AI's challenge of understanding the world. *Science* **382**, eadm8175 (2023).
28. Li, K. et al. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *11th International Conference on Learning Representations* (2023).
29. Lupyan, G., Rahman, R. A., Boroditsky, L. & Clark, A. Effects of language on visual perception. *Trends Cogn. Sci.* **24**, 930–944 (2020).
30. Suffill, E., van Paridon, J. & Lupyan, G. Verbal labels increase conceptual alignment. In *Joint Conference on Language Evolution* (2022).
31. Lake, B. M. & Murphy, G. L. Word meaning in minds and machines. *Psychological Rev.* **130**, 401 (2023).
32. Chemero, A. LLMs differ from human cognition because they are not embodied. *Nat. Hum. Behav.* **7**, 1828–1829 (2023).
33. Warstadt, A. & Bowman, S. R. in *Algebraic Structures in Natural Language* (eds. Lappin, S. & Bernardy, J.-P.) (CRC, 2022).
34. Warriner, A. B., Kuperman, V. & Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013).
35. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
36. Connell, L., Lynott, D. & Banks, B. Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. *Philos. Trans. R. Soc. B* **373**, 20170143 (2018).
37. Niedenthal, P. M., Winkielman, P., Mondillon, L. & Vermeulen, N. Embodiment of emotion concepts. *J. Personal. Soc. Psychol.* **96**, 1120 (2009).
38. Zhong, Y., Huang, C. -R. & Ahrens, K. In *Embodied Grounding of Concreteness/Abstractness: a Sensory-Perceptual Account of Concrete and Abstract Concepts in Mandarin Chinese* (eds. Dong, M. et al.) 72–83 (Springer International, 2022).
39. Connell, L. & Lynott, D. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition* **125**, 452–465 (2012).
40. Sakreida, K. et al. Are abstract action words embodied? An fMRI investigation at the interface between language and motor cognition. *Front. Hum. Neurosci.* **7**, 125 (2013).
41. Fini, C., Era, V., Da Rold, F., Candidi, M. & Borghi, A. M. Abstract concepts in interaction: the need of others when guessing abstract concepts smooths dyadic motor interactions. *R. Soc. Open Sci.* **8**, 201205 (2021).
42. Shapira, N. et al. Clever hans or neural theory of mind? Stress testing social reasoning in large language models. In *Proc. 18th Conference of the European Chapter of the Association for Computational Linguistics* Vol. 1 (eds Graham, Y. & Purver, M.) 2257–2273 (ACL, 2024).
43. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
44. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
45. OpenAI. Gpt-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
46. Team, G. et al. Gemini: a family of highly capable multimodal models. Preprint at https://doi.org/10.48550/arXiv.2312.11805 (2023).
47. OpenAI. Introducing ChatGPT. *ChatGPT* https://openai.com/index/chatgpt/ (2022).
48. Anil, R. et al. Palm 2 technical report. Preprint at https://doi.org/10.48550/arXiv.2305.10403 (2023).
49. Kosslyn, S. M., Ganis, G. & Thompson, W. L. Neural foundations of imagery. *Nat. Rev. Neurosci.* **2**, 635–642 (2001).
50. Pearson, J. The human imagination: The cognitive neuroscience of visual mental imagery. *Nat. Rev. Neurosci.* **20**, 624–634 (2019).
51. Ma, X., Gao, L. & Xu, Q. Tomchallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proc. 27th Conference on Computational Natural Language Learning (CoNLL)* 15–26 (ACL, 2023).
52. Cortese, M. J. & Fugett, A. Imageability ratings for 3,000 monosyllabic words. *Behav. Res. Methods, Instrum., Computers* **36**, 384–387 (2004).
53. Amsel, B. D., Urbach, T. P. & Kutas, M. Perceptual and motor attribute ratings for 559 object concepts. *Behav. Res. Methods* **44**, 1028–1041 (2012).
54. Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504–511 (2024).
55. Bisk, Y. et al. Piqa: Reasoning about physical commonsense in natural language. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 34, 7432–7439 (2020).
56. Kemmerer, D., Rudrauf, D., Manzel, K. & Tranel, D. Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. *Cortex* **48**, 826–848 (2012).
57. Peelen, M. V., He, C., Han, Z., Caramazza, A. & Bi, Y. Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. *J. Neurosci.* **34**, 163–170 (2014).
58. Noppeney, U., Friston, K. J. & Price, C. J. Effects of visual deprivation on the organization of the semantic system. *Brain* **126**, 1620–1627 (2003).
59. Amedi, A., Jacobson, G., Hendler, T., Malach, R. & Zohary, E. Convergence of visual and tactile shape processing in the human lateral occipital complex. *Cereb. cortex* **12**, 1202–1212 (2002).
60. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
61. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Rev.* **104**, 211 (1997).
62. Mikolov, T., Yih, W.-t. & Zweig, G. Linguistic regularities in continuous space word representations. In *Proc. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 746–751 (ACL, 2013).
63. Radford, A. et al. Robust speech recognition via large-scale weak supervision. In *Proc. 40th International Conference on Machine Learning (ICML'23)* (2023).
64. Brohan, A. et al. Rt-2: Vision–language–action models transfer web knowledge to robotic control. In *7th Annual Conference on Robot Learning* (2023).
65. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
66. Elman, J. L. *Rethinking Innateness: A Connectionist Perspective on Development* (MIT Press, 1996).

67. Monaco, E. et al. Embodiment of action-related language in the native and a late foreign language—an FMRI-study. *Brain Lang.* **244**, 105312 (2023).
68. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
69. Liu, A. et al. DeepSeek-V3 Technical Report. Preprint at https://arxiv.org/abs/2412.19437 (2025).
70. Yu, C. & Smith, L. B. The social origins of sustained attention in one-year-old human infants. *Curr. Biol.* **26**, 1235–1240 (2016).
71. Balota, D. A., Pilotti, M. & Cortese, M. J. Subjective frequency estimates for 2,938 monosyllabic words. *Mem. Cognition* **29**, 639–647 (2001).
72. Nastase, S. A. et al. Attention selectively reshapes the geometry of distributed semantic representation. *Cereb. Cortex* **27**, 4277–4291 (2017).
73. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Routledge, 1988).
74. Bonett, D. G. & Wright, T. A. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika* **65**, 23–28 (2000).
75. Myers, L. & Sirois, M. J. Spearman correlation coefficients, differences between. In *Encyclopedia of Statistical Sciences* (eds. Kotz, S. et al.) (Wiley Online Library, 2006).
76. Lynott, D., Connell, L., Brysbaert, M., Brand, J. & Carney, J. Lancaster sensorimotor strength norms: online interactive tool; https://embodiedcognitionlab.shinyapps.io/sensorimotor_norms/ (Lancaster University, 2025).
77. Flaticon terms of use. *Flaticon* https://www.flaticon.com/legal (2024).

Acknowledgements

This research was supported by a grant from the Hong Kong Research Grants Council (project no. PolyU15610322) and the Sin Wai Kin Fundation (P.L.), the Basque Government through the BERC 2022-2025 programme and the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation (grant no. CEX2020-001010/AEI/10.13039/501100011033 to Q.X.) and the Research Postgraduate Scholarships from the Hong Kong Polytechnic University (Y.P.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We express our sincere thanks to V. Sloutsky, V. Valian, J. Magnuson and S. Prasada, as well as to all members of the Brain, Language, and Computation Lab, for their invaluable feedback. Finally, we thank S. Sereno and J. Taylor for sharing the trial-level dataset of the Glasgow Norms.

Author contributions

Q.X., Y.P. and P.L. conceived the project and designed the analyses. Q.X., Y.P. and M.W. collected the data and conducted the analyses. P.L., S.A.N. and M.C. supervised the project. All authors wrote the manuscript and provided critical feedback.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02203-8.

Correspondence and requests for materials should be addressed to Qihui Xu or Ping Li.

Peer review information *Nature Human Behaviour* thanks Laura Bechtold, Yanchao Bi, Anna Borghi, Dermot Lynott and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Qihui Xu; Ping Li

Last updated by author(s): Mar 10, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Data were obtained from the gpt-3.5-turbo-0301 and gpt-4 (collected between May 28 and June 11, 2023) from OpenAI API, and PaLM2 and Gemini-1.0-pro from Google API. All codes are publicly available at https://osf.io/kguwd/

In addition, we developed an analysis pipeline that enables researchers to examine their models of interest. The pipeline and associated resources are also accessible via GitHub: https://github.com/qxu1994/LLM_grounding

Data analysis

Data collection and analysis were done through jupyter notebook and R, with multiple python libraries such as pandas, numpy, scipy, and sklearn. All codes for data collection and analysis are available https://osf.io/kguwd/. Detailed environment information, including software and package versions, is provided below and also within each jupyter notebook and R notebook file. In addition, we developed an analysis pipeline that enables researchers to examine their models of interest. The pipeline and associated resources are also accessible via GitHub: https://github.com/qxu1994/LLM_grounding

For the Lancaster accession and RSA analysis codes:
Python Version: 3.10.0
Pandas Version: 2.2.3
Scipy Version: 1.13.1
Seaborn Version: 0.13.2
Matplotlib Version: 3.9.2
Sklearn (scikit-learn) Version: 1.5.2
Statsmodels Version: 0.14.4

For the rest codes:
Python version: 3.11.7 (main, Dec 15 2023, 12:09:56) [Clang 14.0.6]

IPython: 8.20.0
google-generativeai: 0.7.2
matplotlib: 3.8.0
numpy: 1.26.4
openai: 1.65.5
pandas: 2.1.4
scikit-learn: 1.2.2
scipy: 1.11.4
seaborn: 0.12.2
statsmodels: 0.14.0
tqdm: 4.65.0

For the analyses using R:
R version 4.4.0 (2024-04-24)

package	*	version	date (UTC)	lib	source
abind		1.4-8	2024-09-12	[1]	CRAN (R 4.4.1)
backports		1.5.0	2024-05-23	[2]	CRAN (R 4.4.0)
bayesplot		1.11.1	2024-02-15	[1]	CRAN (R 4.4.0)
bayestestR	*	0.15.0	2024-10-17	[1]	CRAN (R 4.4.1)
boot		1.3-30	2024-02-26	[2]	CRAN (R 4.4.0)
bridgesampling		1.1-2	2021-04-16	[1]	CRAN (R 4.4.0)
brms	*	2.22.0	2024-09-23	[1]	CRAN (R 4.4.1)
Brobdingnag		1.2-9	2022-10-19	[1]	CRAN (R 4.4.0)
broom		1.0.7	2024-09-26	[1]	CRAN (R 4.4.1)
cachem		1.1.0	2024-05-16	[2]	CRAN (R 4.4.0)
car		3.1-3	2024-09-27	[1]	CRAN (R 4.4.1)
carData		3.0-5	2022-01-06	[1]	CRAN (R 4.4.0)
checkmate		2.3.1	2023-12-04	[1]	CRAN (R 4.4.0)
cli		3.6.2	2023-12-11	[2]	CRAN (R 4.4.0)
coda		0.19-4.1	2024-01-31	[1]	CRAN (R 4.4.0)
codetools		0.2-20	2024-03-31	[2]	CRAN (R 4.4.0)
colorspace		2.1-0	2023-01-23	[2]	CRAN (R 4.4.0)
datawizard		0.13.0	2024-10-05	[1]	CRAN (R 4.4.1)
devtools	*	2.4.5	2022-10-11	[1]	CRAN (R 4.4.0)
digest		0.6.35	2024-03-11	[2]	CRAN (R 4.4.0)
distributional		0.5.0	2024-09-17	[1]	CRAN (R 4.4.1)
dplyr		1.1.4	2023-11-17	[2]	CRAN (R 4.4.0)
effectsize		0.8.9	2024-07-03	[1]	CRAN (R 4.4.0)
ellipsis		0.3.2	2021-04-29	[1]	CRAN (R 4.4.1)
emmeans	*	1.10.5	2024-10-14	[1]	CRAN (R 4.4.1)
estimability		1.5.1	2024-05-12	[1]	CRAN (R 4.4.0)
evaluate		1.0.3	2025-01-10	[1]	CRAN (R 4.4.1)
fansi		1.0.6	2023-12-08	[2]	CRAN (R 4.4.0)
fastmap		1.2.0	2024-05-15	[2]	CRAN (R 4.4.0)
Formula		1.2-5	2023-02-24	[1]	CRAN (R 4.4.0)
fs		1.6.4	2024-04-25	[2]	CRAN (R 4.4.0)
generics		0.1.3	2022-07-05	[2]	CRAN (R 4.4.0)
ggplot2	*	3.5.1	2024-04-23	[1]	CRAN (R 4.4.0)
glue		1.7.0	2024-01-09	[2]	CRAN (R 4.4.0)
gridExtra		2.3	2017-09-09	[1]	CRAN (R 4.4.0)
gtable		0.3.5	2024-04-22	[2]	CRAN (R 4.4.0)
htmltools		0.5.8.1	2024-04-04	[2]	CRAN (R 4.4.0)
htmlwidgets		1.6.4	2023-12-06	[1]	CRAN (R 4.4.0)
httpuv		1.6.15	2024-03-26	[1]	CRAN (R 4.4.0)
inline		0.3.19	2021-05-31	[1]	CRAN (R 4.4.0)
insight		0.20.5	2024-10-02	[1]	CRAN (R 4.4.1)
knitr		1.47	2024-05-29	[2]	CRAN (R 4.4.0)
later		1.3.2	2023-12-06	[1]	CRAN (R 4.4.0)
lattice		0.22-6	2024-03-20	[2]	CRAN (R 4.4.0)
lifecycle		1.0.4	2023-11-07	[2]	CRAN (R 4.4.0)
lme4	*	1.1-35.5	2024-07-03	[1]	CRAN (R 4.4.0)
lmerTest	*	3.1-3	2020-10-23	[1]	CRAN (R 4.4.0)
loo		2.8.0	2024-07-03	[1]	CRAN (R 4.4.0)
magrittr		2.0.3	2022-03-30	[2]	CRAN (R 4.4.0)
MASS		7.3-60.2	2024-04-24	[2]	local
Matrix	*	1.7-0	2024-03-22	[2]	CRAN (R 4.4.0)
matrixStats		1.4.1	2024-09-08	[1]	CRAN (R 4.4.1)
memoise		2.0.1	2021-11-26	[2]	CRAN (R 4.4.0)
mime		0.12	2021-09-28	[2]	CRAN (R 4.4.0)
miniUI		0.1.1.1	2018-05-18	[1]	CRAN (R 4.4.0)
minqa		1.2.8	2024-08-17	[1]	CRAN (R 4.4.0)
mnormt		2.1.1	2022-09-26	[1]	CRAN (R 4.4.0)
munsell		0.5.1	2024-04-01	[2]	CRAN (R 4.4.0)

```
mvtnorm        1.3-1     2024-09-03 [1] CRAN (R 4.4.1)
nlme           3.1-164   2023-11-27 [2] CRAN (R 4.4.0)
nloptr         2.1.1     2024-06-25 [1] CRAN (R 4.4.0)
numDeriv       2016.8-1.1 2019-06-06 [2] CRAN (R 4.4.0)
parameters     0.23.0    2024-10-18 [1] CRAN (R 4.4.1)
pillar         1.9.0     2023-03-22 [2] CRAN (R 4.4.0)
pkgbuild       1.4.5     2024-10-28 [1] CRAN (R 4.4.1)
pkgconfig      2.0.3     2019-09-22 [2] CRAN (R 4.4.0)
pkgload        1.4.0     2024-06-28 [1] CRAN (R 4.4.0)
posterior      1.6.0     2024-07-03 [1] CRAN (R 4.4.0)
profvis        0.4.0     2024-09-20 [1] CRAN (R 4.4.1)
promises       1.3.0     2024-04-05 [1] CRAN (R 4.4.0)
psych        * 2.4.6.26  2024-06-27 [1] CRAN (R 4.4.0)
purrr          1.0.2     2023-08-10 [2] CRAN (R 4.4.0)
QuickJSR       1.4.0     2024-10-01 [1] CRAN (R 4.4.1)
R6             2.5.1     2021-08-19 [2] CRAN (R 4.4.0)
Rcpp         * 1.0.13    2024-07-17 [1] CRAN (R 4.4.0)
RcppParallel   5.1.9     2024-08-19 [1] CRAN (R 4.4.1)
remotes        2.5.0     2024-03-17 [1] CRAN (R 4.4.1)
rlang          1.1.4     2024-06-04 [2] CRAN (R 4.4.0)
rmarkdown      2.27      2024-05-17 [2] CRAN (R 4.4.0)
rstan          2.32.6    2024-03-05 [1] CRAN (R 4.4.0)
rstantools     2.4.0     2024-01-31 [1] CRAN (R 4.4.0)
rstatix        0.7.2     2023-02-01 [1] CRAN (R 4.4.0)
rstudioapi     0.16.0    2024-03-24 [2] CRAN (R 4.4.0)
scales         1.3.0     2023-11-28 [2] CRAN (R 4.4.0)
sessioninfo    1.2.3     2025-02-05 [1] CRAN (R 4.4.1)
shiny          1.9.1     2024-08-01 [1] CRAN (R 4.4.0)
StanHeaders    2.32.10   2024-07-15 [1] CRAN (R 4.4.0)
stringi        1.8.4     2024-05-06 [2] CRAN (R 4.4.0)
stringr        1.5.1     2023-11-14 [2] CRAN (R 4.4.0)
tensorA        0.36.2.1  2023-12-13 [1] CRAN (R 4.4.0)
tibble         3.2.1     2023-03-20 [2] CRAN (R 4.4.0)
tidyr          1.3.1     2024-01-24 [2] CRAN (R 4.4.0)
tidyselect     1.2.1     2024-03-11 [2] CRAN (R 4.4.0)
urlchecker     1.0.1     2021-11-30 [1] CRAN (R 4.4.1)
usethis      * 3.1.0     2024-11-26 [1] CRAN (R 4.4.1)
utf8           1.2.4     2023-10-22 [2] CRAN (R 4.4.0)
vctrs          0.6.5     2023-12-01 [2] CRAN (R 4.4.0)
wesanderson  * 0.3.7     2023-10-31 [1] CRAN (R 4.4.0)
withr          3.0.2     2024-10-28 [1] CRAN (R 4.4.1)
xfun           0.45      2024-06-16 [2] CRAN (R 4.4.0)
xtable         1.8-4     2019-04-21 [1] CRAN (R 4.4.0)
yaml           2.3.8     2023-12-11 [2] CRAN (R 4.4.0)
```

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data obtained from ChatGPTs and Google LLMs are publicly available on https://osf.io/kguwd/. The human dataset of the Glasgow norms is from [1], with word-level data accessible at https://doi.org/10.3758/s13428-018-1099-3[32]. The corresponding trial-level data was kindly provided by Sara Sereno and Jack Taylor. The Lancaster norms is from [2], and data including both word-level and trial-level can be found at https://embodiedcognitionlab.shinyapps.io/sensorimotor_norms/ [74]. The validation norms datasets are openly available via the following links: the datasets of valence, arousal, and dominance at https://link.springer.com/article/10.3758/s13428-012-0314-x#SecESM1[33], the imageability norms at https://link.springer.com/article/10.3758/BF03195585#SecESM1[51], the concreteness norms at https://link.springer.com/article/10.3758/s13428-013-0403-5#MOESM1[12], and the perceptual strength norms at https://link.springer.com/article/10.3758/s13428-012-0215-z#SecESM1[52]

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Our paper involves two human datasets from previously published papers [1, 2]. According to [1], among 829 human participants, 599 were females and 230 were males in terms of gender. The paper did not indicate whether sex and/or gender of participants was determined based on self-report or assigned.

	According to [2], among 3,500 human participants, 1,644 were females, 1,823 were males, 12 participants preferred not to indicate their gender and the gender information was missing for 21 participants.
Reporting on race, ethnicity, or other socially relevant groupings	According to [1], the participants were native English speakers from the University of Glasgow community. The paper did not report other socially relevant groupings such as race and ethnicity. According to [2], the participants had English as their first language. The paper did not report other socially relevant groupings such as race and ethnicity.
Population characteristics	According to [1], human participants ranged in age from 16 to 73 years, with a mean of 21.7 years (SD = 7.4). The average age was 21.5 years (SD = 7.6) for female participants and 22.3 years (SD = 6.9) for male participants. According to [2], the average age of all participants was 34.9 years (SD = 10.3).
Recruitment	According to [1], human participants were recruited randomly through an advertisement link on the University of Glasgow's Psychology Department homepage. They received either £6 per hour as compensation or course credit for their participation. According to [2], the participants were recruited via MTurk and only experienced MTurk users who had completed over 100 tasks (i.e., MTurk HITS > 100) with high-quality performance (i.e., > 97% HIT approval) were admitted to the study to ensure data quality. Participants were compensated US$2.75 per completed perceptual strength item list and US$2.25 per completed action strength item list.
Ethics oversight	According to [1], the study adhered to the ethical guidelines and protocols set by the British Psychological Society. According to [2], ethics approval for the study was granted by Lancaster University Research Ethics Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study employs a cross-sectional quantitative design, gathering data from both ChatGPTs and Google LLMs and compared them with ratings generated by humans from the Glasgow and the Lancaster norms. The model prompt and design for LLMs were standardized to match the instructions given to human subjects, maintaining consistency with human-subject data collection. To ensure LLMs' validity as cognitive models, we adopted standard validation techniques from human-subject research for dimensions with notable human-model agreement.
Research sample	We collected data from ChatGPT models - the gpt-3.5-turbo-0301 and gpt-4 (collected between May 28 and June 11, 2023) from the OpenAI API, and Google LLMs - PaLM2 and Gemini-pro-1.0 from the Google API. The temperature parameter was set to 0, following recommendations in (Binz & Schulz, 2023) and (Kosinsky, 2023), to ensure deterministic responses. The maximum token length was set to the upper limits permitted - 2,048 tokens for GPT-3.5, GPT-4, and Gemini, and 1,024 tokens for PaLM, This is a decision informed by the need to capture complete responses without truncation. To enhance the reliability of our results, we implemented four rounds of testing for each model. This approach allowed us to cross-verify the consistency of the outputs across multiple iterations. These models were chosen due to their recognition as state-of-the-art in the field of large language models (LLMs). The significance of using LLMs is highlighted by their ability to a) facilitate the examination of how different modalities (e.g., text language, image, audio, etc.) of input influence learning processes, thereby shedding light on the intricacies of language and cognition, and b) offer avenues for research that transcend the constraints inherent in human-based studies. For comparative analysis with the models' performance, two pre-existing human-generated datasets were utilized: the Glasgow and Lancaster norms. These include: 1. "The Glasgow Norms: Ratings of 5,500 Words on Nine Scales" by Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno, published in Behavior Research Methods, Volume 51, pages 1258–1270, in the year 2019. 2. "The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words" by Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney, also published in Behavior Research Methods, Volume 52, pages 1271–1291, in 2020. Additionally, validation was carried out using several established human-generated datasets, which include: 1. "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas" by Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert, published in Behavior Research Methods, Volume 45, pages 1191–1207, in 2013. 2. "Imageability Ratings for 3,000 Monosyllabic Words" by Michael J. Cortese and April Fugett, featured in Behavior Research Methods, Instruments, & Computers, Volume 36(3), pages 384–387, in 2004. 3. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas" by Marc Brysbaert, Amy Beth Warriner, and

Victor Kuperman, published in Behavior Research Methods, Volume 46, pages 904–911, in 2014.

Sampling strategy

Data collection from the ChatGPT and Google LLM models did not involve a predetermined sample size calculation. Nonetheless, in alignment with prior research, the temperature parameter was set to 0, as recommended in (Binz & Schulz, 2023; Kosinsky, 2023), to guarantee deterministic and consistent responses devoid of random fluctuations. The consistency of the model's responses across multiple runs was assessed, with the agreement metrics detailed in the supplementary material.

Sampling procedure:

For the Glasgow dataset, we used convenience sampling by including all 5,553 words and followed the human-subject sampling procedure [1]. The words were divided into 40 lists: eight lists contained 101 words each, while the remaining 32 lists contained 150 words each. The models rated all words in a list for one dimension before proceeding to the next, with the order of words within each dimension and the order of dimensions in each testing round randomized.

For the Lancaster dataset[2], we extracted all the overlapping words (4,442 words) between the Lancaster norms and the Glasgow norms to ensure uniform comparison. We adhered to the Lancaster norms when creating each testing list: 1) we checked the recognizability of all the 4442 words (more than 95% of words were considered as 'known' by 95% human subjects) before creating testing lists; 2) we extracted the concreteness ratings of all the 4,442 words from [3] and implemented a quantile split based on these concreteness ratings; 3) we randomly extracted 12 words from each quantile and formed 48 testing words for each testing list; 4) we appended the five calibration words in the beginning of each testing list and randomly inserted five control words into each list; 5) For each testing list, the order of the testing words and the order of the sensory and motor dimensions were randomized.

For the validation norms, we again used convenience sampling, by selecting all words that overlapped with those in the Glasgow dataset when validating non-sensorimotor dimensions and with those in the Lancaster dataset when validating sensorimotor dimensions.

Data collection

Data were acquired from the OpenAI API and the Google API, facilitated through Python scripts. The prompt and design for ChatGPT were standardized to align with the guidelines provided to human subjects, thereby ensuring consistency with the methodology used in human-subject data collection. No one else was present besides the researcher. Since data was collected from language models, the researcher was aware of the study hypothesis but did not bias the models' responses.

Timing

Data collection for the Glasgow/Lancaster norms from GPT-3.5 took place from March 25 to June 16th, 2023. For GPT-4, this data was gathered between May 28 and June 11, 2023. Validation norm data for both GPT-3.5 and GPT-4 was collected over June 13th and 14th, 2023. For Google LLMs, this data was collected between March 6th and April 6th, 2024. Validation norm data for Google LLMs was collected between April 25 and May 5, 2024

Data exclusions

In occasional instances, LLMs classified certain words as "unknown," adhering to the guidelines set in the prompt and consistent with protocols from previous human-subject data collections. Typically, these words are ones that may contravene content policies. In line with the practices described in Scott et al. (2019) and Lynott et al. (2020), such data points (i.e., scores from individual runs) were excluded from the data analyses. In addition, PaLM2 failed to return ratings for several lists of words for the sensory dimensions, resulting in ratings for only 2,481 words being collected due to this issue.

Non-participation

No subjects declined participation.

Randomization

This study is correlational, comparing four LLMs—GPT-3.5, GPT-4, PaLM, and Gemini—with human data from previously published studies. These models were not assigned to different groups; all underwent the same testing materials and procedures. While they vary in features such as training size and model architecture, these differences do not impact our evaluation of their similarity to human data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.