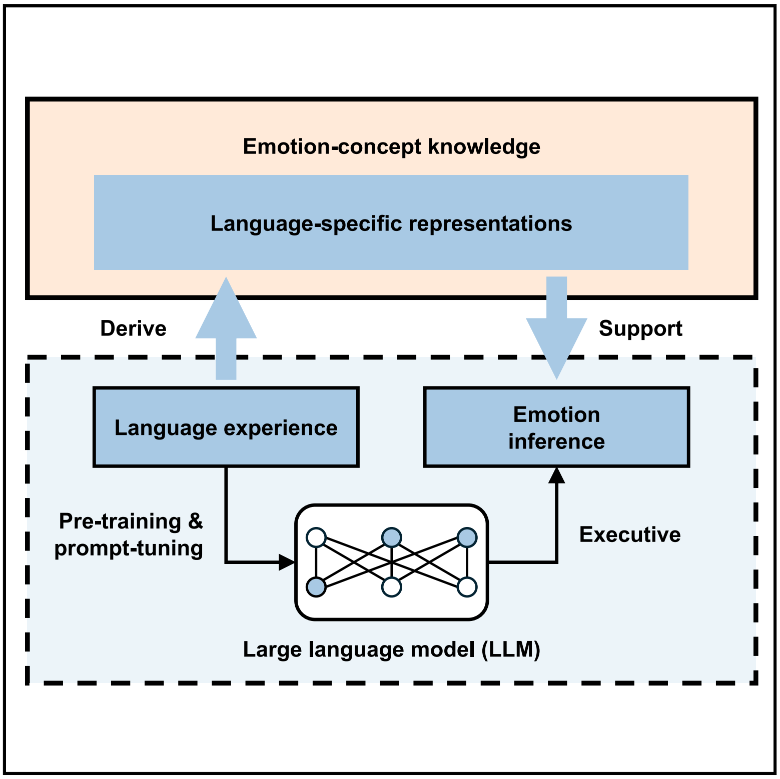
iScience

Article

Language-specific representation of emotion-concept knowledge causally supports emotion inference

Graphical abstract


Emotion-concept knowledge
Language-specific representations
Derive
Support
Language experience
Emotion inference
Pre-training & prompt-tuning
Executive
Large language model (LLM)

Highlights

- Attributes of human emotion concept are represented by distinct LLM neuron sets
- Manipulation of attribute-specific neurons leads to reduced emotion inference
- Attribute-induced reductions are related to their importance in human mental space

Authors

Ming Li (黎明), Yusheng Su (苏裕胜), Hsiu-Yuan Huang (黄绣媛), ..., Kristen A. Lindquist, Zhiyuan Liu (刘知远), Dan Zhang (张丹)

Correspondence

liuzy@tsinghua.edu.cn (Z.L.),
dzhang@tsinghua.edu.cn (D.Z.)

In brief

Artificial intelligence; Emotion in artificial intelligence; Psychology

Li et al., 2024, iScience *27*, 111401
December 20, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111401

CellPress

iScience

Cell Press
OPEN ACCESS

Article

Language-specific representation of emotion-concept knowledge causally supports emotion inference

Ming Li (黎明),[1,2,8] Yusheng Su (苏裕胜),[3,8] Hsiu-Yuan Huang (黄绣媛),[4] Jiali Cheng (成家立),[5] Xin Hu (胡鑫),[6] Xinmiao Zhang (张新淼),[1,2] Huadong Wang (汪华东),[3] Yujia Qin (秦禹嘉),[3] Xiaozhi Wang (王晓智),[3] Kristen A. Lindquist,[7] Zhiyuan Liu (刘知远),[3,*] and Dan Zhang (张丹)[1,2,9,*]

[1]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China
[2]Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[5]Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, MA 01854, USA
[6]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA 15260, USA
[7]Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC 27599, USA
[8]These authors contributed equally
[9]Lead contact
*Correspondence: liuzy@tsinghua.edu.cn (Z.L.), dzhang@tsinghua.edu.cn (D.Z.)
https://doi.org/10.1016/j.isci.2024.111401

SUMMARY

Humans no doubt use language to communicate about their emotional experiences, but does language in turn help humans understand emotions, or is language just a vehicle of communication? This study used a form of artificial intelligence (AI) known as large language models (LLMs) to assess whether language-based representations of emotion causally contribute to the AI's ability to generate inferences about the emotional meaning of novel situations. Fourteen attributes of human emotion concept representation were found to be represented by the LLM's distinct artificial neuron populations. By manipulating these attribute-related neurons, we in turn demonstrated the role of emotion concept knowledge in generative emotion inference. The attribute-specific performance deterioration was related to the importance of different attributes in human mental space. Our findings provide a proof-in-concept that even an LLM can learn about emotions in the absence of sensory-motor representations and highlight the contribution of language-derived emotion-concept knowledge for emotion inference.

INTRODUCTION

At the end of William Shakespeare's *Hamlet*, Horatio looks at Hamlet and says, "Now cracks a noble heart. Good night, sweet prince,/And flights of angels sing thee to thy rest." Although Horatio's facial-bodily expression and tone of voice are imperceptible from this line, the reader can still infer Horatio's grief and admiration. Humans have long been accustomed to communicating individual mental experiences through language.[1,2] However, language's role in inferring others' emotions remains a matter of debate.

This question implicates an ongoing debate regarding the relationship between language and discrete emotion distinctions.[3] Among various perspectives, the traditional ones hold that language processing has no or little effect on the emotion experienced in a given situation because human emotions are considered as categories "embodied" in concrete sensory-motor experiences generated by discrete emotional mechanisms within the brain and body.[4,5] In this view, language primarily plays a communicative role in conveying the meaning of an emotional experience after the fact.[6,7]

In contrast, the constructivist account[8,9] suggests that language experience is involved in shaping the conceptual boundaries of discrete emotions, thereby allowing language-derived knowledge to play an essential role in dictating which emotions are experienced in which contexts. In this view, emotion concepts represent and help differentiate the abstract feature space that comprises emotional experiences. As individuals acquire emotion concepts during early experience, these concepts prospectively warp future experiences of feature space. Similar considerations on color cognition, for illustration, show that congenitally blind people understand color space reasonably well (i.e., which colors are more vs. less similar to one another) solely as a product of learning from other human's linguistic descriptions of the sensorimotor features of color space.[10,11] Suppose language also plays a constitutive role in emotion conceptualization. In that case, human language should be a sufficient, though not the only, way to learn and infer the rich meaning of discrete emotions.[12]

Developmental research supports this perspective by showing that language experience, as a source of emotion-concept knowledge,[13,14] may contribute to children's development of increasingly discrete emotional experiences.[15–17] Increasing behavioral studies further revealed that individual[18,19] and cultural[20,21] differences in emotion-concept knowledge correspond to differences in the sensorimotor representations of emotional facial behaviors. These findings provide correlational evidence for the functional importance of language in emotion conceptualization.

Despite the promising progress, causal evidence for the role of language-grounded knowledge in emotion inference is currently lacking.[22] To this end, researchers have explored manipulations of concept-knowledge representations in the human mind. Through priming techniques,[23] studies have demonstrated that presenting emotion concept words (e.g., ''anger'') before behavioral tasks requiring participants to draw emotional inferences about the meaning of facial behaviors shapes their subsequent perceptions.[24–26] The mechanism of priming is to influence the human mind's access to emotion concepts by pre-activating relevant knowledge representations using priming cues. However, manipulating the access does not directly manipulate knowledge representation *per se*,[27,28] leaving the evidence circumstantial.

A more direct approach involves investigating the behavioral consequences of neurological disorders[29,30] or stimulation[31] in brain regions potentially associated with abstract concept-knowledge representation, i.e., semantic memory.[32] For instance, patients with semantic dementia, characterized by lesions in the anterior temporal lobe (ATL), the ''hub'' of semantic memory,[33] are unable to make inferences about the discrete categorical meaning of facial behaviors.[29] However, the ATL is but one brain region involved in the representation of emotion concepts and is also functionally connected with areas responsible for high-order sensory processing,[34,35] which may support the integration of sensory-grounded modality-specific information.[11,36] Due to the limited understanding of ATL and the difficulty of excluding sensory experience, the extent to which emotion inference relies on non-sensory language-derived knowledge remains uncertain.

Considering the divergence between different theoretical claims, the controversy may be alternatively formulated as whether, without access to sensory-motor representations, discrete emotion concepts can still be learned and inferred through language. Recent developments in large language models (LLMs)[37,38] provide a valuable opportunity to verify this hypothesis by showing the purely language-based representation of emotion-concept knowledge and its causal support for emotion inference.

The investigation of language-derived knowledge in LLMs rests on the assumption that semantically relevant attributes of concepts are reflected in the pattern of linguistic symbol use.[39,40] Thus, text-based computing offers the opportunity to mine linguistic-cultural phenomena for their psychological meanings.[41,42] Many nascent studies have shown that, in the absence of human annotation, LLMs can learn various domains of human knowledge, exclusively on the basis of human language use.[43–45] For instance, recent evidence shows that LLMs can generate moral judgments akin to those made by humans, solely based on learning of human language use.[46]

LLMs are not only useful as concept proofs about the ability of AI to represent complex aspects of human psychology based on human language use alone. LLMs can also be used experimentally to understand how knowledge is applied to novel judgments. The knowledge obtained from pre-training is stored in LLMs and can be selectively activated for different downstream tasks. In this form, the pre-trained LLMs may serve as ideal ''subjects'' who learn about the world exclusively through human language, and we can then inspect their use of emotion-concept knowledge in tasks where the LLM must make an inference about the meaning of a novel emotional situation.

Critically, LLMs can be manipulated more easily than the human brain to understand the relationship between language-specific knowledge representation and emotion inference. Specifically, the artificial neurons in LLMs can be selectively manipulated for their functional relevance to specific concept attributes. In turn, the role of neurons associated with certain content in LLMs can then be causally assessed vis a vis their role in facilitating the LLM's performance on inference tasks.[47] This practice resembles the neural stimulation techniques used in neuroscience research[48–50] but with a higher level of precision (e.g., at the level of single artificial neuron) than is available in human subjects.

Considering the computational principles of language processing shared by LLMs and humans,[51] findings from LLMs have the potential to shed light on the language-based mechanism underlying human emotion inference. This stance of analyzing computational models from a human-like perspective[40,52] has been recognized as beneficial to explore the functional emergence of human cognitive abilities and to address questions that are difficult to answer through human studies alone.[46,53–55]

In the present study, we aimed to explore the language-specific representation of emotion-concept knowledge and its support for emotion inference by LLMs. To justify this possibility, the model we chose needed to be mature enough to perform emotion-related tasks while untuned by user feedback and also provide open parameters for manipulation experiments. We utilized RoBERTa-base[56] in our experiments as the base model, which is a typical LLM that is pre-trained by filling in randomly masked parts of a massive corpus. Since the goal of pre-training is to reproduce as much as possible real human language use, we hypothesize that this LLM can learn human emotion-concept knowledge from the pre-training corpus.

Then the LLM was instructed to perform 27 emotion inference tasks, each requiring it to infer a certain emotion from the same dataset (STAR Methods). To better elicit the LLM's performance, instead of manually designing questioning templates (e.g., ''Does the following sentence express the emotion of grief:''), we trained 27 emotion-specific task prompts using the training set, i.e., prompt tuning.[57] For example, to instruct the LLM to infer remorse, the input would be a combination of the remorse-specific prompt (an optimized questioning template), the text to be inferred, and a placeholder that accepts ''yes'' or ''no'' as the output. The LLM's parameters are frozen during training the task prompts, and then both LLM's parameters and task prompts are immutable for testing.

不可改变的

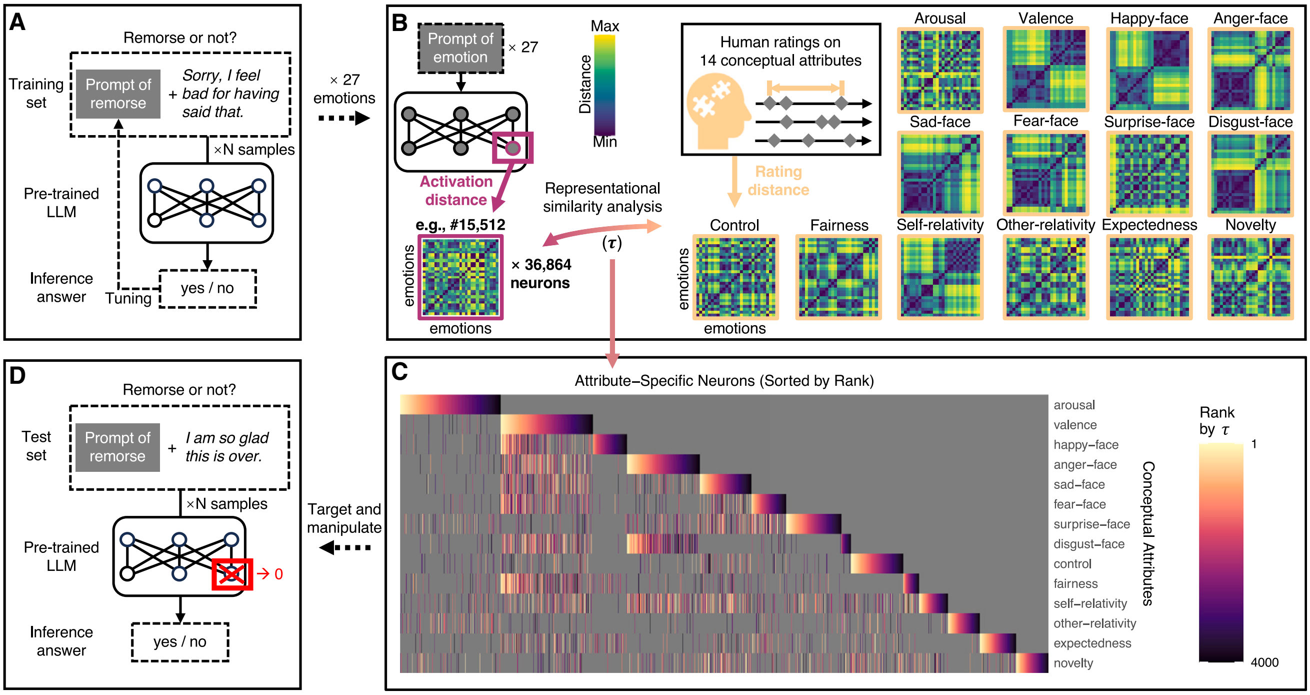


Figure 1. Procedure of manipulating the language-specific representations of emotion-concept knowledge during emotion inference tasks
Illustration of (A) training 27 emotion-specific prompts to stimulate the pre-trained large language model (LLM) for emotion inference tasks on the training set. Illustration of (B) obtaining and validating the LLM-based knowledge representation by searchlight representational similarity analysis (RSA)[58] between every artificial neuron's activation in response to 27 prompts and the human ratings on 14 conceptual attributes. According to the rank of Kendall's *tau* values, the top *N* neurons most relevant to 14 conceptual attributes were (C) targeted and (D) manipulated during emotion inference tasks on the test set. The example of $N = 4{,}000$ was shown in (C). Euclidean distance data in (B) are calculated based on mean neuron activations or mean item scores. See details in STAR Methods and full results of different *N* in Figure S5.

Since the LLM's behavior (knowing which emotion to infer) is determined by task prompts, artificial neuron activations elicited by the prompt itself were expected to represent the corresponding concept knowledge. We then input only emotion-specific prompts to the LLM, without any *concrete stimuli*, to obtain its neuron representation at task level (STAR Methods and Figure 1B). Since no sample text was input in this operation, our *task abstraction* approach prevents the possible contamination of knowledge representation by stimulus keywords. To validate the representational content of LLM-based emotion-concept knowledge, we conducted behavioral experiments to obtain human ratings on 14 attributes of emotion concept from existing literature, including core affects, prototypical expressions, and antecedent appraisals. We compared these ratings with the LLM's representations from a higher-level functional perspective through representational similarity analysis (RSA).[58] Subsequently, guided by the representational space of the human ratings, we could locate and manipulate the artificial neurons in the LLM relevant to a conceptual attribute (e.g., valence) of emotions to investigate their causal contributions for emotion inference tasks. We further explored the association between the language-based contribution of different conceptual attributes and their importance in human mental space, which could provide evidence for a deeper understanding of the constitution and inference of emotion concepts.

RESULTS

LLM infers emotions based on shared conceptualization

After optimizing 27 emotion-specific prompts to infer the corresponding emotion on the training set, we reported the average accuracies of task prompts over different random seeds by evaluating the test set (Table S1, see also Figure 2A). The average accuracy for each of the 27 emotion inference tasks varied from 68.04% (realization) to 96.43% (gratitude).

The LLM's inference accuracy was positively related to rater agreement on dataset annotations. Pearson's *r*(25) was 0.797 with $p < 0.001$, suggesting that the more agreement human raters had on a particular type of emotional scenarios, the more accurate the LLM's inference is (see also Table S1; Figure S1).

Language-specific representations of emotion-concept knowledge

As the LLM's neuron activations in response to an emotion-specific prompt solely without text samples were considered to represent the corresponding emotion concepts, we investigated their representational content using RSA. This technique aims to evaluate the second-order similarity between the representational dissimilarity matrices (RDMs) of every single artificial neuron's activation and the RDMs of 14 conceptual attributes of emotion, which were obtained from human rating experiments (STAR Methods). See Figure 1B for illustration.

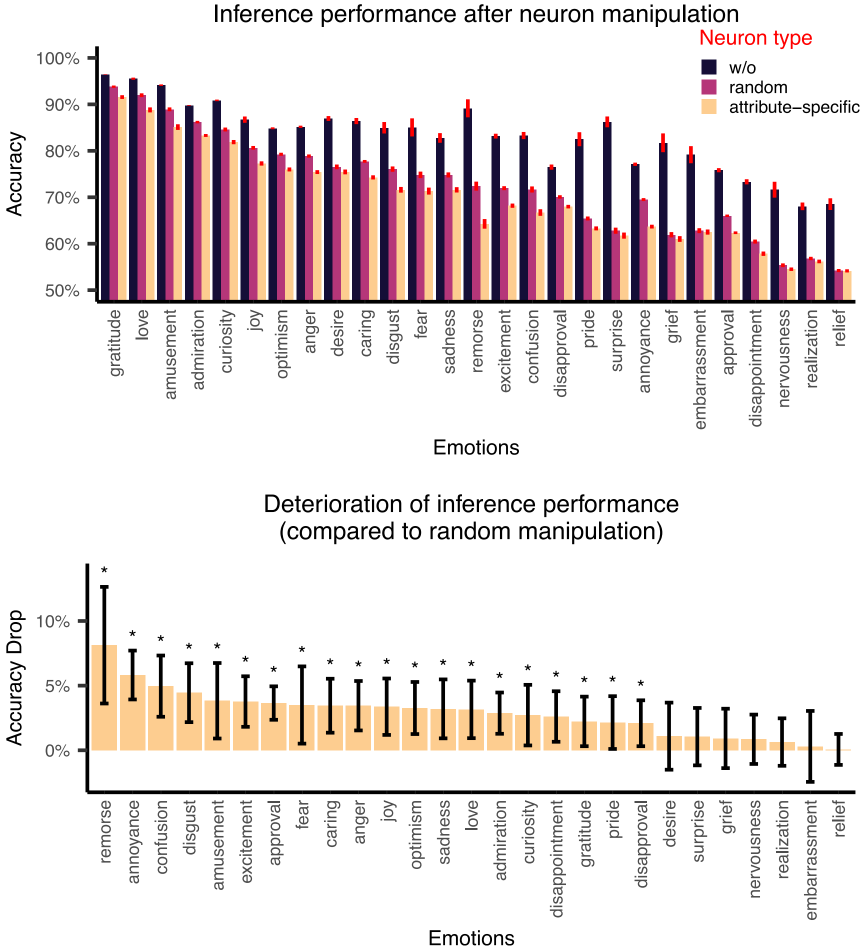


Figure 2. Results of manipulating the language-specific representations of emotion-concept knowledge during emotion inference tasks

The inference accuracy was evaluated (A) with manipulating N attribute-specific neurons, N random neurons, or without manipulation. Data are represented as mean ± SEM over task's random seeds (and manipulated attributes). The causal contribution of emotion-concept knowledge was indicated by (B) the accuracy drops due to manipulated neuron type (attribute-specific vs. random). One-tailed paired t-test was conducted on accuracy drop among task's random seeds and manipulated attributes. Asterisks indicate $p < 0.05$ with false discovery rate (FDR)[59] corrected across all inference tasks and numbers of manipulated neurons. Error bars indicate 95% CIs of mean. The example of $N = 4{,}000$ was shown in (A and B). See details in STAR Methods and full results of different N in Figures S6 and S7.

The second-order similarities exhibited that each of the 14 conceptual attributes was significantly related to a subset of artificial neurons, which were distributed in all layers of the LLM rather than concentrated in specific layers. The results are shown in Figures S5 and S6, false discovery rate (FDR) corrected $q < 0.01$, one-tailed sign-rank test.

According to the rank of relatedness (Kendall's *tau*) between artificial neurons and conceptual attributes, the most relevant neurons for different attributes were less overlapping. The example of the top 4,000 attribute-specific neurons is shown in Figure 1C; see Figure S7 for results of different top N.

Emotion-concept knowledge causally contributes to emotion inference

The reliance of emotion inference on language-derived emotion-concept knowledge was then revealed by manipulating attribute-specific neurons while inferring 27 emotions (Figure 1D). See STAR Methods for the details of the manipulation experiment. Compared to the original accuracy without manipulation, we found a drop in the accuracy of emotion inference on LLM with selective manipulation (Figures 2A and S8).

This deterioration of inference performance still held when compared to randomly manipulating the same number of neurons, suggesting the unique causal contribution of emotion-concept knowledge representations *per se* (Figure S9). Significance was determined by a one-tailed paired t-test on accuracy drop averaged across random seeds and conceptual attributes, FDR corrected $p < 0.05$. The most prominent performance deterioration arose when manipulating the top 4,000 attribute-specific neurons, shown in Figure 2B.

The possible differences in the knowledge contribution for inferring different emotions were examined by testing the heterogeneity of 27 tasks' performance deterioration. Hartigan's dip test showed no evidence of significant heterogeneity regarding the reliance on emotion-concept knowledge for different tasks, with minimum $p = 0.127$ for every conceptual attribute and every number of manipulated neurons (Table S2).

Importance in mental space predicts language-based contribution of different attributes

To further explain the contributions of different conceptual attributes to emotion inference (Figure 3A), we also estimated the weight of different knowledge in the human mental representation of emotion concepts (Figure 3B) and compared the evidence from two parts. See STAR Methods for the details of human experiment and related analysis.

For the LLM with access only to natural language, 12 of the 14 conceptual attributes contribute significantly to emotion inference under specific numbers of manipulated neurons in varying degrees, except for "self-relativity" and "disgust-face" (Figure 3A; for details, see Figure S10). Significance was determined by a one-tailed paired t-test on accuracy drop averaged across random seeds and inference tasks, FDR corrected $p < 0.05$.

For humans with both normal language and sensory functions, their mental representations reflect 12 of the 14 conceptual attributes, except for "arousal" and "other-relativity". Two-tailed

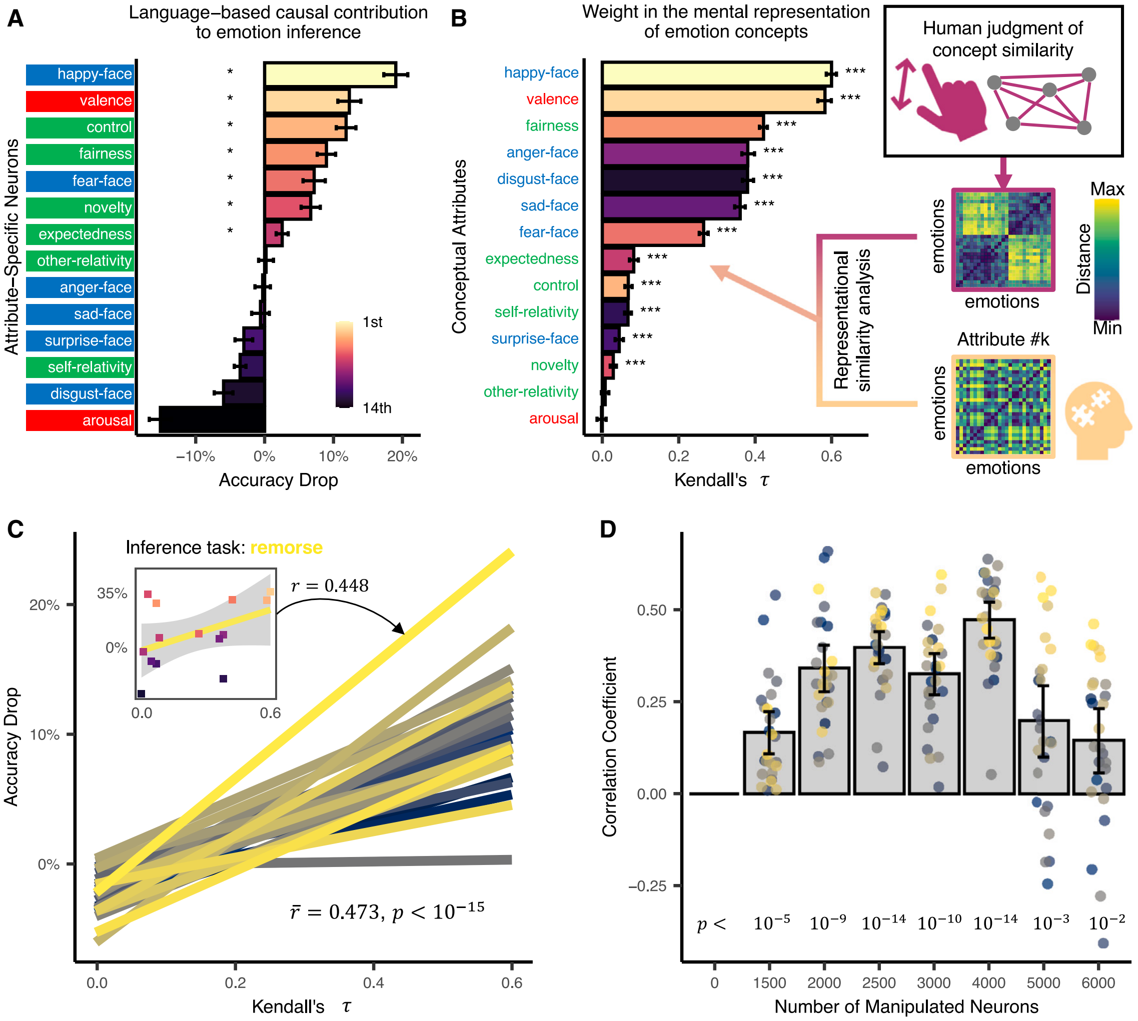


Figure 3. Comparison between the language-based knowledge contribution to emotion inference and knowledge weight in the human mental representation of emotion concepts

(A) The language-based knowledge contribution to emotion inference was indicated by mean accuracy drop (vs. random manipulation), with one-tailed paired *t*-test across inference tasks and task's random seeds. The asterisks indicate $p < 0.05$ with FDR corrected across all conceptual attributes and numbers of manipulated neurons. Here, the example of manipulating 4,000 neurons was shown (see Figure S10 for full results).

(B) Knowledge weight in the human mental representation of emotion concepts was estimated by RSA's *tau*, and the human mental representation of emotion-concepts ($N = 61$) was obtained by the concept similarity judgment experiment. The asterisks indicate FDR corrected bootstrap-based $p < 0.001$ across all conceptual attributes. See STAR Methods for details. For both (A and B), colors of attribute labels indicate “core affects (red)”, “prototypical expressions (blue)”, and “antecedent appraisals (green)”.

(C and D) The conceptual attributes' weights in human mental representation were then correlated to language-based contributions on (subplot in C) arbitrary emotion inference task when (C) manipulating the top 4,000 attribute-specific neurons and (D) manipulating a different number of neurons. Lines in (C) are fitted by linear regression. The significances in (C) and (D) were determined by one-tailed *t*-test on Fisher's transformed correlation coefficients. All error bars indicate 95% CIs of mean value.

signed-rank test with bootstrap sampling, FDR corrected $p < 0.001$ across all attributes (Figure 3B).

We further demonstrated that the language-based knowledge contribution to emotion inference was not independent, but significantly related to knowledge weight in human mental representation. The strongest correlation arose when manipulating the top 4,000 attribute-specific neurons in the LLM, with Fisher-based average $r = 0.473$, $t(26) = -16.714$, $SD = 0.030$, $p < 10^{-15}$ (Figure 3C). In addition, as the number of manipulated neurons increased from 1,500 to 6,000, an inverted-U shaped relationship was observed (Figure 3D), suggesting that there may be “floor effects” and “ceiling effects”[60] to manifest the reliance of emotion inference on emotion-concept knowledge in the LLM.

DISCUSSION

In the present study, we adopted a human-like perspective to investigate the representation of emotion-concept knowledge in an LLM in response to the theoretical debates about the relationship between language and discrete emotion distinctions. Our results provide a proof-in-concept that, even without any access to sensory-motor representations, various attributes of emotion concepts can still be abstracted from language experience and causally support emotion inferences. This computational evidence makes no demand for the experiential feelings once thought essential for discrete emotion differentiation, though, it correlated well with behavioral evidence of the importance of these conceptual attributes.

The findings that the LLM's neuron activations effectively represented emotion-concept knowledge extend our understanding of language-accessible human knowledge to the emotion domain (Figure S5). While previous studies have preliminarily shown the association between language learning and emotion conceptualization,[13,61] we further demonstrated that attributes of emotion concept could derive from the sensory-independent language experience, i.e., the statistical regularities among the linguistic symbols. This computational evidence points to a valuable yet understudied hypothesis: humans may also be able to learn emotion concepts directly from everyday language use.[16,62]

The consistency of LLM's emotion inference performance with human raters' agreement on emotion annotations further indicates that what is reflected in the large-scale language corpus is a shared understanding among individuals (Table S1 and Figure S1). Intriguingly, the understanding includes not only cognitive appraisal attributes, but also more experiential attributes, i.e., the associated core affects and similarity with prototypical facial expressions (STAR Methods). The latter may suggest that even human judgments measured with visual stimuli (faces) contain meaningful information that can be cross-modal, such as a social cue of threat reflected in "fear" eyes.[63] Given the computational nature of LLMs,[64] it would be interesting to elucidate further how individual/cultural differences in abstract symbol use influence emotional development.[14,65]

Moreover, the LLM's knowledge-related artificial neurons were distributed across all layers (Figure S6), possibly implying the involvement of both low-level (e.g., phrasal features) and high-level (e.g., syntactic and semantic features) linguistic regularities for emotion conceptualization.[66] The distributions of the artificial neurons corresponding to different conceptual attributes also tended to be distinct (Figures 1C and S7), suggesting possibly unique contributions of these attributes. These results may inspire further exploration of the neural mechanisms underlying language-derived knowledge representations about emotion concepts.

Most importantly, our manipulation of attribute-specific neurons in the LLM (Figures 1D, 2A and 2B) examined the functional validity of the language-specific representations of emotion-concept knowledge in emotion inference. This result could contribute to a central and ongoing debate in emotion science about the nature of human emotion categories,[67–70] i.e., whether language only superficially conveys emotional experiences after the fact, or whether emotion experiences are constructed in part via abstract conceptual category knowledge acquired via language.[71] Whereas previous lesion studies revealed the necessity of semantic memory for discrete emotion differentiation,[29] we used the LLM to demonstrate further that language-specific knowledge representations are sufficient, at least in principle, to differentiate discrete emotions from meaningful contexts (Figures 2B and 3A; see Figures S8–S10 for whole results). By pointing out the weak heterogeneity in the knowledge contribution to 27 emotions (Table S2), we suggest a unifying mechanism for LLMs and possibly also for the human brain to infer various emotions. Our view is reinforced by the recent neuroimaging findings that one broad ensemble containing multiple brain networks represents a range of emotions[72] rather than distinct emotions consistently and specifically activating local brain regions.[73]

It is worth noting that the language-based emotion inference mechanism is not exclusive of sensory-motor processes, nor of other semantic processes integrated with sensory-motor experiences, such as meaning making[12] and prototype matching.[74] Instead, our comparison between the language-based knowledge contribution and the knowledge weight in human mental space suggests that language has limitations in supporting emotion inference. For example, albeit the conceptual similarity with "disgust-face" can be reflected by both the LLM's artificial neurons (Figures S5 and S6; see also Figure 1C) and human mental representation (Figure 3B), its language-specific representation *did not* effectively contribute to emotion inference (Figure S7). Based on this study, it is hard to elucidate whether the language-specific representation of specific conceptual attributes ("self-relativity" and "disgust-face") needs to be combined with other experiential information to be functional. However, the correlation between the language-based knowledge contributions and the knowledge weights in human mental representation (Figures 3C and 3D) illustrates that human conceptualization of discrete emotions is inextricably linked to the functionality of the language-derived knowledge.

Since LLMs and the human brain have been suggested to share similar computational principles for language,[51,75,76] LLMs can serve as a potential reference in the future to help us understand the actual, rather than hypothetical, language-dependent algorithms that the human brain relies on to infer emotions. One promising direction is to integrate the computational models to uncover key empirical evidence in comparisons between typical and atypical populations, such as those with congenital perceptual deficits and/or late language exposure. Future research could also use LLMs to investigate how different semantic processes[11,36] drive the supramodal representation of emotions in the brain. For example, growing evidence suggested that the brain can convergently process and integrate emotional cues across modalities (e.g., facial expressions and prosody) and represent their conceptual meaning in amodal areas.[77] Suppose the activity of these amodal areas during emotion perception fits with the LLM's hidden state values. In that case, possible neural mechanisms of language-dependent semantic processing involved in making emotional meanings from sensory input can be revealed.

On the other hand, the fact that a machine can learn rich and meaningful knowledge of human emotions through natural language, without "true feelings", also points in a humane and affordable way to develop artificial agents with social functions.[78]

However, since the introduction of "affective computing" concept,[79] the role of domain-general language experience in machine's ability to recognize emotions through different behavioral patterns has not been fully recognized. For example, there is some competing evidence from computer vision suggesting that emotion recognition, at least in vision, relies on domain-specific rather than broad experience, e.g., faces and scenes.[52,80,81] By migrating our methodology to multimodal large models, it is expected to reveal the uniqueness of different data modalities and how they should be integrated for facilitating machine emotional intelligence.

In conclusion, recent advances in LLMs provide a rare opportunity to test the independent role of language in conceptualizing and inferring discrete emotions. The fact that even an LLM can learn the rich meaning of emotions solely through human language will have broad implications for the fields of human psychology and artificial intelligence. How two basic functions, language and sensory motor, interact with each other in emotion-related cognitive and neural mechanisms should be the focus of future research.

Limitations of the study

Since our aim was to investigate whether emotion concepts derived from human language are sufficient to help differentiate emotional meaning from novel scenarios, the use of the LLM completely precluded access to sensory-motor representations. In this context, our results cannot answer, and are not intended to answer, whether emotional feelings are rooted in language use.

Furthermore, the generalizability of the findings may be limited as a single type of LLM (RoBERTa-base) and one specific dataset (GoEmotions) were employed. While this encoder-only LLM was chosen for its balance of performance, stability, and openness, this does not entail that LLMs trained with other strategies (e.g., the decoder-only LlaMa family[82]) will provide evidence of the same strength. Likewise, conducting similar experiments on datasets in other cultural or linguistic contexts (e.g., non-English language with different emotion conceptualizations[83]) would help to better understand the universals and variations of the hypothesis tested. In addition, while human behavioral results could be influenced by other factors, such as gender, age, etc., the potential impact of these factors is believed to be limited, as our findings are largely consistent with the results from the original papers that employed these paradigms.[14,21,84,85] Nevertheless, it is necessary to address these factors in future studies to have a more complete overview of their possible influences.

Another limitation of this study is that we only tested the role of language-derived emotion knowledge in emotion inference but not in other tasks (which requires annotations other than emotion labels). In other words, it remains an open question to what extent such language-based knowledge constructs are shared by emotionally relevant (domain-specific) and emotionally irrelevant (domain-general) cognitive processes.[8,16] To further explore this direction, it is necessary for psychologists and computer scientists to collaborate on developing benchmark datasets with annotations rich in human cognitive domains.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Dan Zhang (dzhang@tsinghua.edu.cn).

Materials availability

The LLM, the task dataset, and the stimulus used in human experiments are publicly available or are fully described.

Data and code availability

- The task dataset for LLM experiments[86] is publicly available at https://doi.org/10.18653/v1/2020.acl-main.372. The raw data associated with this study, including human behavioral scores and neuron activations, have been deposited at https://github.com/thunlp/Model_Emotion. Accession numbers are listed in the key resources table.
- The code to replicate this study, including the scripts to obtain artificial neuron activation, to perform RSA, and to conduct neuron manipulation experiments, is available at https://github.com/thunlp/Model_Emotion.
- All other materials, including the open-source base LLM and the stimuli for human behavioral experiments, are listed in the key resources table.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (T2341003, 62236004), the National Key Research and Development Program of China (No. 2020AAA0106500), the National Science Foundation of China (NSFC) and the German Research Foundation (DFG) in project Cross-modal Learning (NSFC 62061136001/DFG TRR-169/C1, C4), the Teaching Reform Project of the Instruction Committee of Psychology in Higher Education by the Ministry of Education of China (20222008), and the Education Innovation Grants, Tsinghua University (DX05_02).

AUTHOR CONTRIBUTIONS

Conceptualization: D.Z. and Z.L.; methodology: Y.S., H.W., Y.Q., and X.W.; formal analysis: M.L., Y.S., X.H., and X.Z.; investigation: M.L., H.-Y.H., and J.C.; writing – original draft preparation: M.L., D.Z., K.A.L., and Y.S.; writing – review and editing: X.H., X.Z., Z.L., H.-Y.H., J.C., H.W., Y.Q., and X.W.; funding acquisition: D.Z. and Z.L.; project administration: M.L. and Y.S.; resources: H.W., Y.Q., and X.W.; software: M.L., Y.S., H.-Y.H., and J.C.; supervision: D.Z.; visualization: M.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Large language model and task dataset
  - ○ Conceptual attributes rating experiments
  - ○ Concept similarity judgment experiment
- METHOD DETAILS
  - ○ Training of emotion-specific task prompts
  - ○ Evaluation of LLM's inference performance
  - ○ Model-based knowledge representation
  - ○ Conceptual attributes rating experiments
  - ○ Artificial neuron manipulation experiment
  - ○ Concept similarity judgment experiment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Searchlight RSA for LLM's representation of emotion-concept knowledge

- ○ Task reliance on emotion-concept knowledge
- ○ Heterogeneity test for the reliance of different tasks
- ○ Language-based knowledge contribution
- ○ RSA for knowledge weight in mental representation of emotion concepts
- ○ Comparison between language-based knowledge contribution and knowledge weight in mental representation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111401.

Received: March 12, 2024
Revised: July 10, 2024
Accepted: November 12, 2024
Published: November 15, 2024

REFERENCES

1. Baumard, N., Huillery, E., Hyafil, A., and Safra, L. (2022). The cultural evolution of love in literary history. Nat. Human Behav. *6*, 506–522. https://doi.org/10.1038/s41562-022-01292-z.
2. Lindquist, K.A., Jackson, J.C., Leshin, J., Satpute, A.B., and Gendron, M. (2022). The cultural evolution of emotion. Nat. Rev. Psychol. *1*, 669–681. https://doi.org/10.1038/s44159-022-00105-4.
3. Satpute, A.B., and Lindquist, K.A. (2021). At the Neural Intersection Between Language and Emotion. Affect. Sci. *2*, 207–220. https://doi.org/10.1007/s42761-021-00032-2.
4. Ekman, P., and Cordaro, D. (2011). What is meant by calling emotions basic. Emotion Review *3*, 364–370. https://doi.org/10.1177/1754073911410740.
5. Tracy, J.L., and Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. Emotion Review *3*, 397–405. https://doi.org/10.1177/1754073911410747.
6. Ekman, P. (1992). An Argument for Basic Emotions. Cognit. Emot. *6*, 169–200. https://doi.org/10.1080/02699939208411068.
7. Keltner, D., and Cordaro, D.T. (2017). Understanding multimodal emotional expressions: Recent advances in basic emotion theory. In Russell J.A.,Fernández-Dols J.-M., editors. The science of facial expression (Oxford University Press), pp. 57–75.
8. Lindquist, K.A. (2013). Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model. Emotion Review *5*, 356–368. https://doi.org/10.1177/1754073913489750.
9. Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cognit. Affect Neurosci. *12*, 1–23. https://doi.org/10.1093/scan/nsw154.
10. Shepard, R.N., and Cooper, L.A. (1992). Representation of Colors in the Blind, Color-Blind, and Normally Sighted. Psychol. Sci. *3*, 97–104. https://doi.org/10.1111/j.1467-9280.1992.tb00006.x.
11. Bi, Y. (2021). Dual coding of knowledge in the human brain. Trends Cognit. Sci. *25*, 883–895. https://doi.org/10.1016/j.tics.2021.07.006.
12. Satpute, A.B., and Lindquist, K.A. (2019). The Default Mode Network's Role in Discrete Emotion. Trends Cognit. Sci. *23*, 851–864. https://doi.org/10.1016/j.tics.2019.07.003.
13. Shablack, H., Becker, M., and Lindquist, K.A. (2020). How do children learn novel emotion words? A study of emotion concept acquisition in preschoolers. J. Exp. Psychol. Gen. *149*, 1537–1553. https://doi.org/10.1037/xge0000727.
14. Nook, E.C., Sasse, S.F., Lambert, H.K., McLaughlin, K.A., and Somerville, L.H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. Nat. Human Behav. *1*, 881–889. https://doi.org/10.1038/s41562-017-0238-7.
15. Streubel, B., Gunzenhauser, C., Grosse, G., and Saalbach, H. (2020). Emotion-specific vocabulary and its contribution to emotion understanding in 4- to 9-year-old children. J. Exp. Child Psychol. *193*, 104790. https://doi.org/10.1016/j.jecp.2019.104790.
16. Hoemann, K., Wu, R., LoBue, V., Oakes, L.M., Xu, F., and Barrett, L.F. (2020). Developing an Understanding of Emotion Categories: Lessons from Objects. Trends Cognit. Sci. *24*, 39–51. https://doi.org/10.1016/j.tics.2019.10.010.
17. Nencheva, M.L., Dusen, H.V., Watson, E., and Lew-Williams, C. (2023). Natural dynamics of caregiver-child affect are linked to communication and children's word knowledge. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/wbsym.
18. Brooks, J.A., and Freeman, J.B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. Nat. Human Behav. *2*, 581–591. https://doi.org/10.1038/s41562-018-0376-6.
19. Hu, X., Wang, F., and Zhang, D. (2022). Similar brains blend emotion in similar ways : Neural representations of individual difference in emotion profiles. Neuroimage *247*, 118819. https://doi.org/10.1016/j.neuroimage.2021.118819.
20. Gendron, M., Roberson, D., van der Vyver, J.M., and Barrett, L.F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. Emotion *14*, 251–262. https://doi.org/10.1037/a0036052.
21. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., and Pollak, S.D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychol. Sci. Publ. Interest *20*, 1–68. https://doi.org/10.1177/1529100619832930.
22. Lindquist, K.A. (2017). The role of language in emotion: existing evidence and future directions. Curr. Opin. Psychol. *17*, 135–139. https://doi.org/10.1016/j.copsyc.2017.07.006.
23. Maxfield, L. (1997). Attention and Semantic Priming: A Review of Prime Task Effects. Conscious. Cognit. *6*, 204–218. https://doi.org/10.1006/ccog.1997.0311.
24. Lindquist, K.A., Satpute, A.B., and Gendron, M. (2015). Does Language Do More Than Communicate Emotion? Curr. Dir. Psychol. Sci. *24*, 99–108. https://doi.org/10.1177/0963721414553440.
25. Nook, E.C., Lindquist, K.A., and Zaki, J. (2015). A new look at emotion perception: Concepts speed and shape facial emotion recognition. Emotion *15*, 569–578. https://doi.org/10.1037/a0039166.
26. Gendron, M., Lindquist, K.A., Barsalou, L., and Barrett, L.F. (2012). Emotion words shape emotion percepts. Emotion *12*, 314–325. https://doi.org/10.1037/a0026007.
27. Firestone, C., and Scholl, B.J. (2014). "Top-Down" Effects Where None Should Be Found: The El Greco Fallacy in Perception Research. Psychol. Sci. *25*, 38–46. https://doi.org/10.1177/0956797613485092.
28. Firestone, C., and Scholl, B.J. (2016). Cognition does not affect perception: Evaluating the evidence for top-down effects. Behav. Brain Sci. *39*, e229–e277. https://doi.org/10.1017/S0140525X15000965.
29. Lindquist, K.A., Gendron, M., Barrett, L.F., and Dickerson, B.C. (2014). Emotion perception, but not affect perception, is impaired with semantic memory loss. Emotion *14*, 375–387. https://doi.org/10.1037/a0035293.
30. Jastorff, J., de Winter, F.L., van den Stock, J., Vandenberghe, R., Giese, M.A., and Vandenbulcke, M. (2016). Functional dissociation between anterior temporal lobe and inferior frontal gyrus in the processing of dynamic body expressions: Insights from behavioral variant frontotemporal dementia. Hum. Brain Mapp. *37*, 4472–4486. https://doi.org/10.1002/hbm.23322.
31. Long, Y., Zhong, M., Aili, R., Zhang, H., Fang, X., and Lu, C. (2023). Transcranial direct current stimulation of the right anterior temporal lobe changes interpersonal neural synchronization and shared mental processes. Brain Stimul. *16*, 28–39. https://doi.org/10.1016/j.brs.2022.12.009.
32. Tulving, E. (1972). Episodic and semantic memory. In Organization of Memory, E. Tulving and W. Donaldson, eds. (Academic Press), pp. 382–402.

33. Patterson, K., Nestor, P.J., and Rogers, T.T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. Nat. Rev. Neurosci. *8*, 976–987. https://doi.org/10.1038/nrn2277.

34. Sabsevitz, D.S., Medler, D.A., Seidenberg, M., and Binder, J.R. (2005). Modulation of the semantic system by word imageability. Neuroimage *27*, 188–200. https://doi.org/10.1016/j.neuroimage.2005.04.012.

35. Mahon, B.Z., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J. Physiol. Paris *102*, 59–70. https://doi.org/10.1016/j.jphysparis.2008.03.004.

36. Popham, S.F., Huth, A.G., Bilenko, N.Y., Deniz, F., Gao, J.S., Nunez-Elizalde, A.O., and Gallant, J.L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. Nat. Neurosci. *24*, 1628–1636. https://doi.org/10.1038/s41593-021-00921-6.

37. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Preprint at arXiv. https://doi.org/10.48550/arXiv.2005.14165.

38. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., and Riedel, S. (2019). Language models as knowledge bases? EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. In Proceedings of the Conference (Association for Computational Linguistics), pp. 2463–2473. https://doi.org/10.18653/v1/d19-1250.

39. Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. Perspect. Psychol. Sci. *14*, 1006–1033. https://doi.org/10.1177/1745691619861372.

40. Grand, G., Blank, I.A., Pereira, F., and Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. Nat. Human Behav. *6*, 975–987. https://doi.org/10.1038/s41562-022-01316-8.

41. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books Team; Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., et al. (2011). Quantitative analysis of culture using millions of digitized books. Science *331*, 176–182. https://doi.org/10.1126/science.1199644.

42. Jackson, J.C., Watts, J., List, J.M., Puryear, C., Drabble, R., and Lindquist, K.A. (2022). From Text to Thought: How Analyzing Language Can Advance Psychological Science. Perspect. Psychol. Sci. *17*, 805–826. https://doi.org/10.1177/17456916211004899.

43. Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Trans. Assoc. Comput. Linguist. *8*, 34–48. https://doi.org/10.1162/tacl_a_00298.

44. Aspillaga, C., Mendoza, M., and Soto, A. (2021). Inspecting the concept knowledge graph encoded by modern language models. Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2984–3000. https://doi.org/10.18653/v1/2021.findings-acl.263.

45. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., and Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat. Mach. Intell. *4*, 258–268. https://doi.org/10.1038/s42256-022-00458-8.

46. Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI Language Models Replace Human Participants? Trends in Cognitive Sciences, 597–600. https://doi.org/10.1016/j.tics.2023.04.008.

47. Wang, X., Wen, K., Zhang, Z., Hou, L., Liu, Z., and Li, J. (2022). Finding Skill Neurons in Pre-trained Transformer-based Language Models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics)), pp. 11132–11152.

48. Walsh, V., and Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. Nat. Rev. Neurosci. *1*, 73–79. https://doi.org/10.1038/35036239.

49. Tong, J., Kong, C., Wang, X., Liu, H., Li, B., and He, Y. (2020). Transcranial direct current stimulation influences bilingual language control mechanism: evidence from cross-frequency coupling. Cogn. Neurodyn. *14*, 203–214. https://doi.org/10.1007/s11571-019-09561-w.

50. Liu, S., He, Y., Guo, D., Liu, X., Hao, X., Hu, P., and Ming, D. (2023). Transcranial alternating current stimulation ameliorates emotional attention through neural oscillations modulation. Cogn. Neurodyn. *17*, 1473–1483. https://doi.org/10.1007/s11571-022-09880-5.

51. Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. *25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4.

52. Zhou, L., Yang, A., Meng, M., and Zhou, K. (2022). Emerged human-like facial expression representation in a deep convolutional neural network. Sci. Adv. *8*, eabj4383–12. https://doi.org/10.1126/sciadv.abj4383.

53. Doerig, A., Sommers, R.P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G.W., Kording, K.P., Konkle, T., van Gerven, M.A.J., Kriegeskorte, N., and Kietzmann, T.C. (2023). The neuroconnectionist research programme. Nat. Rev. Neurosci. *24*, 431–450. https://doi.org/10.1038/s41583-023-00705-w.

54. Frank, M.C. (2023). Baby steps in evaluating the capacities of large language models. Nat. Rev. Psychol. *2*, 451–452. https://doi.org/10.1038/s44159-023-00211-x.

55. Kanwisher, N., Khosla, M., and Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. Trends Neurosci. *46*, 240–254. https://doi.org/10.1016/j.tins.2022.12.008.

56. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.11692.

57. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv. *55*, 1–35. https://doi.org/10.1145/3560815.

58. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4–28. https://doi.org/10.3389/neuro.06.004.2008.

59. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188. https://doi.org/10.1214/aos/1013699998.

60. Lim, C.R., Harris, K., Dawson, J., Beard, D.J., Fitzpatrick, R., and Price, A.J. (2015). Floor and ceiling effects in the OHS: An analysis of the NHS PROMs data set. BMJ Open *5*, e007765. https://doi.org/10.1136/bmjopen-2015-007765.

61. Hoemann, K., Hartley, L., Watanabe, A., Solana Leon, E., Katsumi, Y., Barrett, L.F., and Quigley, K.S. (2021). The N400 indexes acquisition of novel emotion concepts via conceptual combination. Psychophysiology *58*, 1–13. https://doi.org/10.1111/psyp.13727.

62. Hoemann, K., Xu, F., and Barrett, L.F. (2019). Emotion Words, Emotion Concepts, and Emotional Development in Children: A Constructionist Hypothesis. Dev. Psychol. *55*, 1830–1849. https://doi.org/10.1037/dev0000686.

63. Jessen, S., and Grossmann, T. (2014). Unconscious discrimination of social cues from eye whites in infants. Proc. Natl. Acad. Sci. USA *111*, 16208–16213. https://doi.org/10.1073/pnas.1411333111.

64. Blank, I.A. (2023). What are large language models supposed to model?. Preprint at arXiv. https://doi.org/10.1016/j.tics.2023.08.006.

65. Camacho, M.C., Nielsen, A.N., Balser, D., Furtado, E., Steinberger, D.C., Fruchtman, L., Culver, J.P., Sylvester, C.M., and Barch, D.M. (2023). Large-scale encoding of emotion concepts becomes increasingly similar between individuals from childhood to adolescence. Nat. Neurosci. *26*, 1256–1266. https://doi.org/10.1038/s41593-023-01358-9.

66. Jawahar, G., Sagot, B., and Seddah, D. (2020). What does BERT learn about the structure of language? In ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Association for Computational Linguistics (ACL)), pp. 3651–3657. https://doi.org/10.18653/v1/p19-1356.

67. Barrett, L.F. (2006). Are Emotions Natural Kinds? Perspect. Psychol. Sci. *1*, 28–58. https://doi.org/10.1111/j.1745-6916.2006.00003.x.

68. Adolphs, R., Mlodinow, L., and Barrett, L.F. (2019). What is an emotion? Curr. Biol. *29*, R1060–R1064. https://doi.org/10.1016/j.cub.2019.09.008.

69. Fridlund, A.J. (2017). The behavioral ecology view of facial displays, 25 years later. In The science of facial expression, J.A. Russell and J.-M. Fernandez-Dols, eds. (Oxford University Press), pp. 77–92. https://doi.org/10.1093/acprof:oso/9780190613501.003.0005.

70. Cowen, A.S., and Keltner, D. (2021). Semantic Space Theory: A Computational Approach to Emotion. Trends in Cognitive Sciences *25*, 124–136. https://doi.org/10.1016/j.tics.2020.11.004.

71. Barrett, L.F., and Westlin, C. (2021). Navigating the science of emotion. In Emotion Measurement (Elsevier), pp. 39–84. https://doi.org/10.1016/b978-0-12-821124-3.00002-8.

72. Horikawa, T., Cowen, A.S., Keltner, D., and Kamitani, Y. (2020). The Neural Representation of Visually Evoked Emotion Is High-Dimensional, Categorical, and Distributed across Transmodal Brain Regions. iScience *23*, 101060. https://doi.org/10.1016/j.isci.2020.101060.

73. Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., and Barrett, L.F. (2012). The brain basis of emotion: A meta-analytic review. Behav. Brain Sci. *35*, 121–143. https://doi.org/10.1017/S0140525X11000446.

74. Binetti, N., Roubtsova, N., Carlisi, C., Cosker, D., Viding, E., and Mareschal, I. (2022). Genetic algorithms reveal profound individual differences in emotion recognition. Proc. Natl. Acad. Sci. USA *119*, e2201380119. https://doi.org/10.1073/pnas.2201380119.

75. Caucheteux, C., and King, J.R. (2022). Brains and algorithms partially converge in natural language processing. Commun. Biol. *5*, 134. https://doi.org/10.1038/s42003-022-03036-1.

76. Tuckute, G., Kanwisher, N., and Fedorenko, E. (2024). Language in Brains, Minds, and Machines. Annu. Rev. Neurosci. *47*, 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142.

77. Schirmer, A., and Adolphs, R. (2017). Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. Trends Cognit. Sci. *21*, 216–228. https://doi.org/10.1016/j.tics.2017.01.001.

78. De Melo, C.M., Gratch, J., Marsella, S., and Pelachaud, C. (2023). Social Functions of Machine Emotional Expressions. Proc. IEEE *111*, 1382–1397. https://doi.org/10.1109/JPROC.2023.3261137.

79. Picard, R.W. (1997). Affective Computing (The MIT Press). https://doi.org/10.1037/E526112012-054.

80. Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Sci. Adv. *8*, eabl8913–12. https://doi.org/10.1126/sciadv.abl8913.

81. Kragel, P.A., Reddan, M.C., LaBar, K.S., and Wager, T.D. (2019). Emotion schemas are embedded in the human visual system. Sci. Adv. *5*, eaaw4358. https://doi.org/10.1126/sciadv.aaw4358.

82. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2307.09288.

83. Jackson, J.C., Watts, J., Henry, T.R., List, J.M., Forkel, R., Mucha, P.J., Greenhill, S.J., Gray, R.D., and Lindquist, K.A. (2019). Emotion semantics show both cultural variation and universal structure. Science *366*, 1517–1522. https://doi.org/10.1126/science.aaw8160.

84. Russell, J.A. (2003). Core Affect and the Psychological Construction of Emotion. Psychol. Rev. *110*, 145–172. https://doi.org/10.1037/0033-295X.110.1.145.

85. Skerry, A.E., Saxe, R., Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., Kersting, K., Kragel, P.A., Reddan, M.C., LaBar, K.S., et al. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. Curr. Biol. *25*, 1945–1954. https://doi.org/10.1016/j.cub.2015.06.009.

86. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372.

87. Lundqvist, D., Flykt, A., and Ohman, A. (1998). The Averaged Karolinska Directed Emotional Faces - AKDEF (CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet). https://doi.org/10.1037/t27732-000.

88. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. *20*, 37–46. https://doi.org/10.1177/001316446002000104.

89. Su, Y., Wang, X., Qin, Y., Chan, C.M., Lin, Y., Wang, H., Wen, K., Liu, Z., Li, P., Li, J., et al. (2022). On Transferability of Prompt Tuning for Natural Language Processing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Association for Computational Linguistics)), pp. 3949–3969. https://doi.org/10.18653/v1/2022.naacl-main.290.

90. Barrett, L.F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. Pers. Soc. Psychol. Rev. *10*, 20–46. https://doi.org/10.1207/s15327957pspr1001_2.

91. Levenson, R.W. (2011). Basic emotion questions. Emotion Review *3*, 379–386. https://doi.org/10.1177/1754073911410743.

92. Du, S., Tao, Y., and Martinez, A.M. (2014). Compound facial expressions of emotion. Proc. Natl. Acad. Sci. USA *111*, E1454–E1462. https://doi.org/10.1073/pnas.1322355111.

93. Scherer, K.R., and Fontaine, J.R.J. (2019). The semantic structure of emotion words across languages is consistent with componential appraisal models of emotion. Cognit. Emot. *33*, 673–682. https://doi.org/10.1080/02699931.2018.1481369.

94. Clore, G., and Ortony, A. (2016). Psychological Construction in the OCC Model of Emotion Gerald. Emot. Rev. *5*, 335–343. https://doi.org/10.1177/1754073913489751.Psychological.

95. Roseman, I.J., Spindel, M.S., and Jose, P.E. (1990). Appraisals of Emotion-Eliciting Events: Testing a Theory of Discrete Emotions. J. Pers. Soc. Psychol. *59*, 899–915. https://doi.org/10.1037/0022-3514.59.5.899.

96. Hartigan, J.A., and Hartigan, P.M. (1985). The dip test of unimodality. Ann. Stat. *13*, 70–84.

97. Freeman, J.B., and Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. Behav. Res. Methods *45*, 83–97. https://doi.org/10.3758/s13428-012-0225-x.

98. Heffner, J., and FeldmanHall, O. (2022). A probabilistic map of emotional experiences during competitive social interactions. Nat. Commun. *13*, 1718–1811. https://doi.org/10.1038/s41467-022-29372-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GoEmotions (LLM task dataset)	Demszky et al., 2020[86]	https://doi.org/10.18653/v1/2020.acl-main.372
AKDEF stimuli set (for conceptual attributes rating experiment)	Lundqvist & Flykt, 1998[87]	https://kdef.se/index.html
Event-appraisal items (for conceptual attributes rating experiment)	Skerry et al., 2015[85]	https://doi.org/10.1016/j.cub.2015.06.009
Raw data associated with this paper	This paper	https://github.com/thunlp/Model_Emotion
Software and algorithms		
RoBERTa-base	Liu et al., 2019[56]	https://doi.org/10.48550/arXiv.1907.11692
Python (version 3.8.0)	Python Software Foundation	RRID:SCR_008394; https://www.python.org/
R (version 4.1.2)	R Foundation	RRID:SCR_001905; http://www.r-project.org/
MATLAB (version 2021a)	MathWorks	RRID:SCR_001622; https://matlab.mathworks.com
Other		
All code used in this paper	This paper	https://github.com/thunlp/Model_Emotion

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Large language model and task dataset

The LLM utilized in this study is RoBERTa-base.[56] The text dataset we employed for emotion inference tasks in this study is a public corpus GoEmotions[86] (https://github.com/google-research/google-research/tree/master/goemotions), which contains 58,009 English Reddit comments, manually labeled with 27 emotions by 82 native English speakers from India. To perform different inference tasks on the same dataset, we used 27 emotion-specific task prompts to instruct the LLM on which emotion should be inferred. For example, to instruct the LLM to infer remorse, the input would be a combination of three parts – i.e., [instance to be inferred] + [prompt of remorse] + [MASK]. The first part is arbitrary text sample taken from the test set representing an emotional scenario that the LLM has not seen before. The second part is a task prompt that instructs the LLM to determine whether the sample expresses remorse or not, which is trained using the training set and the corresponding remorse labels. The last part is a special placeholder that will be filled in by the LLM with "yes" or "no" as its answer. Following the authors of the dataset, we divided the dataset into training (80%), development (10%), and test (10%) sets. We did not use the development set in any subsequent operation because its proposed purpose was incompatible with this study.

Conceptual attributes rating experiments

For each of three attribute sets (core affects, prototypical expressions, and antecedent appraisals), we obtained human ratings through an independent online experiment on the Prolific website. For core affects, thirty participants from Australia ($n = 2$), Canada ($n = 5$), Ireland ($n = 2$), South Africa ($n = 5$), United Kingdom ($n = 15$), and United States ($n = 1$) were recruited (mean age = 33 years, 15 female). For prototypical expressions, thirty participants from Australia ($n = 1$), Canada ($n = 5$), Ireland ($n = 3$), Hungary ($n = 1$), Netherlands ($n = 1$), South Africa ($n = 11$), Spain ($n = 1$), United Kingdom ($n = 6$), United States ($n = 1$) were recruited (mean age = 31 years, 17 female). For antecedent appraisals, two hundred ninety-nine participants from Australia ($n = 7$), Canada ($n = 13$), France ($n = 1$), Greece ($n = 1$), Ireland ($n = 5$), Israel ($n = 1$), Italy ($n = 3$), Japan ($n = 1$), Nea Zealand ($n = 3$), Poland ($n = 2$), Portugal ($n = 4$), South Africa ($n = 36$), Spain ($n = 1$), Sweden ($n = 1$), Switzerland ($n = 1$), United Kingdom ($n = 210$), United States ($n = 10$) were recruited (mean age = 37 years, 148 female). All participants were native English speakers without language related disorders. Subjects in all experiments were required to complete all trials except for the appraisal rating experiment in which subjects were randomly assigned to one of the 27 emotions. In addition to the adopted participants reported above, we excluded 5, 6, and 30 subjects from the three experiments due to failure of the attention check, respectively. The Institutional Review Board at the Department of Psychology, Tsinghua University, approved all experimental procedures. All participants gave their informed consent. All participants gave their informed consent.

Concept similarity judgment experiment

Another sixty-one English-speaking participants from Australia ($n = 4$), Canada ($n = 5$), France ($n = 1$), Hungary ($n = 1$), New Zealand ($n = 3$), Slovenia ($n = 1$), South Africa ($n = 3$), United Kingdom ($n = 28$), United States ($n = 15$) were recruited (mean age = 36 years, 30 female) from the Prolific website and asked to complete the experiment online. All participants were native English speakers without language related disorders. Subjects were required to complete all trials and no subject was excluded. Institutional Review Board at the Department of Psychology, Tsinghua University, approved the experimental procedures. All participants gave their informed consent.

METHOD DETAILS

Training of emotion-specific task prompts

To find valid prompts for each task, we optimized only the prompts (some pseudo tokens that function as questioning templates) while freezing all parameters of the LLM to generate the appropriate answer (yes/no) for the input text (Figure 1A).

Formally, $\mathcal{M}$ was RoBERTa. Given an input text with n tokens $\mathrm{X} = \{\mathrm{w}_1, \mathrm{w}_2, \cdots, \mathrm{w}_n\}$, RoBERTa first converted them into input embeddings $\mathbf{X_e} \in \mathrm{R}^{\mathrm{n} \times \mathrm{d}}$, where d was the dimension of the embedding space. We pre-pended l randomly initialized trainable tokens $\mathbf{P_e} \in \mathrm{R}^{\mathrm{l} \times \mathrm{d}}$ before the input matrix $\mathrm{X_e}$, and formed the modified input embeddings $[\mathrm{P_e}, \mathrm{X_e}] \in \mathrm{R}^{(\mathrm{l}+\mathrm{n}) \times \mathrm{d}}$. A special [MASK] token was additionally pre-pended before the prompts, which would output the probability of label tokens. The objective (O) was to maximize the likelihood of the desired output y:

$$O = P_{\mathcal{M}}([MASK] = y|[\boldsymbol{P_e}, \boldsymbol{X_e}]).$$

During the prompt tuning, we only optimized the trainable tokens (P_e) while freezing the whole parameters of a RoBERTa ($\mathcal{M}$) to maximize the above objective. Freezing parameters of the LLM prevented task information from changing the LLM's learned knowledge.

To obtain the corresponding prompt of each emotion on RoBERTa, we re-framed the 27-class emotion dataset of GoEmotions into 27 emotion inference tasks. For instance, for the emotion "remorse", if a text belonged to the category "remorse", then we re-labeled the text with y = "yes"; otherwise, y = "no". In this way, we obtained the new training data for each emotion. During training, we set the prompt length to $l = 100$ and the prompt dimension to $d = 768$. After conducting prompt tuning individually for each emotion inference task, we obtained all prompts $\{P_e^c \in R^{100 \times 768} | c \in \mathcal{C}\}$, where $\mathcal{C}$ was the set of 27 discrete emotions. The function of the prompts is to instruct the LLM to perform corresponding tasks, but their format need neither contain emotion words nor be human-readable. In other words, the task prompts and the output answers are related to specific emotion concepts, but do not contain their lexical forms.

In order to avoid statistical bias, for each emotion inference task, we trained prompts 12 times with 12 random seeds; all of these 12 prompts have been evaluated on the test set, respectively.

Evaluation of LLM's inference performance

All task prompts with all random seeds have been evaluated on the test set, respectively (Table S1). Since the LLM is pre-trained on large-scale human language corpora, it should be more capable of inferring emotions with more shared conceptualization. We then estimated the agreement of the dataset raters for each emotion via Cohen's Kappa,[88] and related it with the LLM's inference performance on the test set (Figure S1).

Model-based knowledge representation

Since these prompts can activate the corresponding LLM task state,[47,89] the LLM's neuron activation values in response to the emotion-specific prompts were considered to represent knowledge about the corresponding emotion concept. Hence, we input the trained prompts for each of the 27 emotions without concatenating any text into the LLM to activate the hidden states values, also known as the LLM's artificial neuron activations (Figure 1B).

In our setting, the values of artificial neurons s were the values of hidden states between the FFN layer in a Transformer. Specifically, we could denote the FFN layer as:

$$\mathrm{FFN}(x) = \mathrm{GELU}\left(xW_1^\top + b_1\right)W_2 + b_2,$$

where $x \in R^d$ was the input embedding, $W_1, W_2 \in R^{d_m \times d}$ were trainable matrices, and b_1, b_2 were bias vectors. The value of artificial neurons was $= xW_1^\top + b_1$.

For each task, we input the sequence, $\{[MASK], P, < s >\}$, into RoBERTa, where P was the emotion-specific prompt, $\langle s < > \rangle$ was the special token indicating the start of an input sentence. Finally, we stacked the values of artificial neurons in all FFN layers of RoBERTa to get the overall neuron activation values AS(P) for each emotion inference task:

$$\mathrm{AS}(P) = [v_1; v_2; \ldots; v_L],$$

where $L = 36,864$ was the total number of artificial neurons. These activation values of 36,864 artificial neurons were collected for further analysis.

Conceptual attributes rating experiments

To measure the content of emotion-concept knowledge, we chose the most representative attributes of emotion concepts in the existing psychological emotion theories, including three sets: 2 core affect attributes, 6 prototypical expression attributes, and 6 antecedent appraisal attributes.

The core affect attributes address how an emotion concept would feel to a person in a general sense.[84,90] The emotions in lexical form were presented randomly for each participant, followed by their literal definition (consistent with the GoEmotions dataset[86]) and a nine-point Likert scale for both attributes. There was text instruction above each rating scale, "To what extent does [*EMOTION*] make you feel... (Valence: 1 = very unpleasant, 5 = neutral, 9 = very pleasant; Arousal: 1 = very calming, 9 = very arousing)". There were no other stimuli or context around.

The prototypical expression attributes reflect the similarity of emotion concepts to six stereotypical emotional faces,[87] conforming to the classical operational definitions of Basic Emotion Theory.[91,92] A nine-point Likert scale was presented to participants, and each participant's order of emotions (with literal definition) and faces were randomized. There was text and image instruction above each rating scale, "To what extent is [*EMOTION*] consistent with the physiological responses shown in the figures: (1 = very inconsistent, 5 = neutral, 9 = very consistent)". The images we used to indicate six prototypical expressions are twelve averaged faces (one male and one female for each prototypical expression) from the AKDEF stimulus set[87] (https://kdef.se/index.html).

As the antecedent appraisal attributes are often associated with event-specific features,[93,94] we instruct participant to recall an event that caused them to feel one of the 27 emotions (randomly assigned) and rate 38 items on the event. In the recall phase, we instructed participants to remember and write down a situation (at least 100 words) in which they felt the given emotion (with literal definition) and then identify the specific event (up to 50 words) in the situation that directly caused that emotion. This procedure avoided involving multiple events, cognitions, and emotions in a single recall.[95] We instructed participants in the next phase to rate 38 items for that specific event in random order. All those items were summarized by a previous study,[85] covering most factors from the appraisal theories of emotion. Before the next processing step, we kept six factors in the 299 events times 38 items matrix as appraisal attributes (see Figures S2 for factor details).

All above conceptual attributes were averaged across repeated ratings as the final attribute scores for each emotion concept. The final scores and their reliabilities are shown in Figures S3 and S4.

Artificial neuron manipulation experiment

To examine the potential support of emotion-concept knowledge representation for emotion inference, we input the trained prompts and the test set of scenarios into LLM to infer emotions. During the inference of 27 emotions, we modified the activation values of attribute-specific neurons to zero (Figure 1D). For each conceptual attribute, the number of manipulated neurons was set uniformly to 1500, 2000, 2500, 3000, 4000, 5000, or 6000. Overall, the selective manipulation operation was repeated 34,020 times, respectively, for 14 conceptual attributes, 27 emotion inference tasks (12 prompts/random seeds per task), and seven levels (the number of manipulated neurons).

To exclude the influence of manipulating neurons *per se*, we randomly select the same number of neurons to manipulate as a control group for every operation. The causal contribution of emotion-concept knowledge was indicated as the difference in accuracy after selective manipulation compared to accuracy after random manipulation, i.e., accuracy drop due to the neuron type (attribute-specific vs. random).

Concept similarity judgment experiment

We adopted a similarity judgment task to measure the human mental representation of the 27 emotion concepts (Figure 3B). Sixty-one English-speaking participants (30 females, mean age = 36 years) were recruited from Prolific and asked to complete the task online. They judged the subjective similarity of 27 emotion concepts (and "neutral" concept) using a 9-point Likert scale (1 = most dissimilar, 9 = most similar) without criteria cues. These 27 emotion and neutral concepts were presented simultaneously on the screen in word form. However, participants judged the similarity of only the two words with black borders each time. There were no response time limits but instructions to participants to respond by first sense when they hesitated.

We retained similarity scores between 27 emotions (351 pairs) and replaced missing values (3 per participant) with the average score across participants. Then, for each participant, these scores were subtracted by 10 to indicate the dissimilarity (ranging from 1 to 9) and used to form an individual representational dissimilarity matrix (RDM),[58] i.e., a 27 by 27 symmetric matrix with a diagonal of 0 to indicate that any emotion is equal to itself. Each RDM reflected one participant's mental representation of emotion concepts.

QUANTIFICATION AND STATISTICAL ANALYSIS

Searchlight RSA for LLM's representation of emotion-concept knowledge

We first built an RDM for each artificial neuron. An RDM is a symmetric matrix (27 emotions by 27 emotions), where the elements are the Euclidean distances of the random-seed-averaged neuron activation in response to emotion-specific prompts. The RDMs for 14 conceptual attributes were also built, respectively, by calculating the Euclidean distances of emotion concepts' final score on that attribute.

We then conducted the one-tailed sign-rank test to indicate the relatedness between the RDM of each artificial neuron and the RDM of each conceptual attribute, using only the lower triangle of RDMs (Figure 1B). Due to many calculations, we did not perform the bootstrap method. Instead, we used false discovery rate (FDR) correction[59] to control multiple comparisons for all neurons and all attributes. The significance of Kendall's *tau* values shows the absolute correspondence of the artificial neurons for each conceptual attribute (Figures S5 and S6), and the rank of Kendall's *tau* values shows the relative correspondences (Figures 1C and S7).

Task reliance on emotion-concept knowledge

For each of the 27 emotion inference tasks at a specific number of manipulated neurons, we estimated its reliance on emotion-concept knowledge by one-tailed paired *t*-test, i.e., accuracy drop averaged across task prompt's random seeds and 14 conceptual attributes. FDR was corrected to determine the significance of accuracy drops. The example of manipulating 4,000 neurons is shown in Figure 2.

Heterogeneity test for the reliance of different tasks

To explain whether there are systematic differences in the reliance on emotion-concept knowledge for inferring different emotions, we then tested the heterogeneity of 27 tasks' reliance on specific conceptual attribute with specific number of manipulated neurons. This analysis was conducted by examining the multimodality (i.e., two or more distinct peaks) of random-seed-averaged accuracy drop via Hartigan's dip,[96] which is a statistical test for exploring whether the data may be from different groups or different distributions.[97,98] A significant result would suggest that inferences on different emotion concepts by knowledge representations are heterogeneous or qualitatively different.

Language-based knowledge contribution

The causal contribution of a specific conceptual attribute to emotion inference was estimated by accuracy drop (vs. random manipulation) averaged across emotion inference tasks and task prompt's random seeds, at a specific number of manipulated neurons. One-tailed paired *t*-test was conducted. FDR was corrected to determine the significance. The example of manipulating 4,000 neurons is shown in Figure 3A.

RSA for knowledge weight in mental representation of emotion concepts

To estimate the weight of specific conceptual attributes in the human mental representation of emotion concepts, we conducted RSA to show how well the RDM of a conceptual attribute fit the RDMs of people's emotion-concept representations (Kendall's *tau*; Figure 3B). The RDM of each of 14 conceptual attributes was related to people's RDMs via the two-tailed signed-rank test, with bootstrap sampling participants and emotions 1,000 times. FDR was corrected to control multiple comparisons across 14 attributes.

Comparison between language-based knowledge contribution and knowledge weight in mental representation

The degrees of accuracy drop for 14 conceptual attributes were then correlated with these conceptual attributes' weights in human mental representations, i.e., Kendall's *tau* values. Considering the weak heterogeneity of knowledge contribution across emotion inference tasks (see Table S2), we treated each emotion inference task as a sample set and obtained a series of Pearson's correlation coefficients to determine significance by the one-tailed *t*-test on Fisher's transformed coefficients (for an illustration, see example in Figure 3C).