

Cortical knowledge structures guide word concept learning

Received: 21 July 2025

Accepted: 21 April 2026

Published online: 13 May 2026

Guangyao Zhang¹, Xiaosha Wang^{2,3,4}, Dingchen Zhang¹, Siwen Xie^{5,6},
Lusha Zhu^{1,2,3,4,7} ✉ & Yanchao Bi^{1,2,3,4,5} ✉

Human word concept learning leverages prior knowledge to generalize from limited exemplars, yet its neural implementation remains unclear. We developed a Neural Bayesian Model (NBM) that incorporates neural representational priors to explain word concept learning. Using fMRI, we first constructed a neural prior space from activity elicited by familiar objects (with novel shapes as controls), and then examined neural responses during novel word concept learning based on these objects. The NBM integrating priors from ventral occipitotemporal cortex predicted both neural representations and behavioral generalization, outperforming control models lacking neural priors. In contrast, hippocampal activity supported learning for novel shapes without benefit from the NBM, consistent with prior-free associative mechanisms. Large language models showed weaker alignment with human generalization patterns. These findings dissociate prior-based cortical inference from hippocampal exemplar-associative learning, providing a neural instantiation of Bayesian concept acquisition and clarifying the interplay between semantic and episodic memory systems.

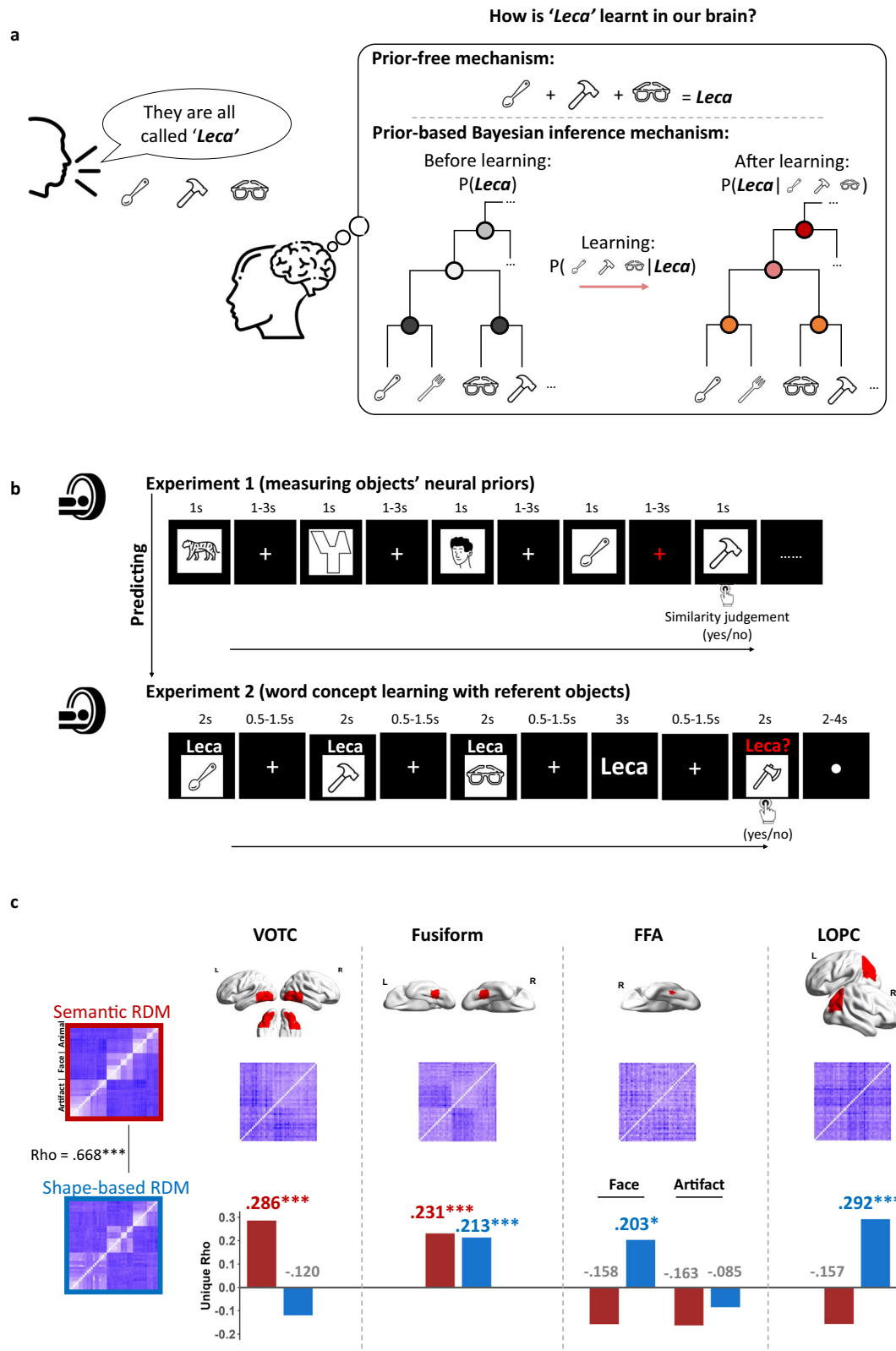
What happens in the human brain when one learns the meaning of an unknown word (e.g., “Leca”) via a reference object, say, a spoon? Remarkably, humans can often infer the meanings of such novel words from just a few exemplars and apply them appropriately to new instances. The ability to learn and generalize word meanings from sparse input is a hallmark of human intelligence. Does it follow a prior-free approach, mapping words directly to corresponding exemplars, or a prior-based approach that leverages existing conceptual knowledge—particularly knowledge about the categorical structures and relationships of exemplars (Fig. 1a)? It has been proposed that such background knowledge, which is likely grounded in our long-term semantic memory, constrains candidate hypotheses about the word meaning and guides generalization, thereby enabling rapid and data-efficient learning of word concepts¹. This prior-based word concept learning is a powerful mechanism for constructing and learning new word concepts^{2–5} and is considered a cornerstone of

human intelligence, as well as a critical challenge for artificial intelligence models^{6–8}.

Despite its theoretical appeal, the neural mechanism by which the knowledge stored in semantic memory supports word concept learning remains elusive. Neuroscience research has implicated classical regions in various forms of concept and category learning. One line of research, derived from associative learning literature, has examined concept learning by focusing on the formation of artificially designed categorization rules or associations, typically minimizing the influence of prior knowledge. These studies have converged to reveal the roles of the hippocampus and ventromedial prefrontal cortex (VMPFC) in learning new (conceptual) categories/associations^{9–12}. In particular, the role of the hippocampus and adjacent medial temporal regions in learning via building new associations between stimuli has been highlighted^{13–18}, suggesting a prior-free learning mechanism. Another line of research has specifically

¹Faculty of Psychology, Beijing Normal University, Beijing, China. ²School of Psychological and Cognitive Sciences, Peking University, Beijing, China. ³IDG/McGovern Institute for Brain Research, Peking University, Beijing, China. ⁴Key Laboratory of Machine Perception (Ministry of Education), Beijing, China.

⁵Institute for Artificial Intelligence, Peking University, Beijing, China. ⁶Yuanpei College, Peking University, Beijing, China. ⁷Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. ✉ e-mail: lushazhu@pku.edu.cn; ybi@pku.edu.cn



examined how prior knowledge influences the learning of novel associations, reporting that the presence and strength of prior knowledge increased activity in the dorsomedial prefrontal cortex (DMPFC) and decreased activity in the hippocampus^{19,20}. These findings resonate with the growing focus on the interaction between semantic memory (i.e., long-term store of knowledge about the world) and episodic learning^{21–23}, but how neural representations in

semantic memory precisely participate in forming new word concepts remains unknown.

In contrast, cognitive developmental research has clarified, at the behavioral level, the computational principles guiding the integration of background semantic knowledge with newly learned exemplars in support of word concept learning. A prominent class of models formalizes this process as Bayesian inference over a structured

Fig. 1 | Schematic overview of the experimental design and analyses. **a** Example illustrating the core question: how a new word concept (e.g., “Leca”) is learned via a few exemplars (e.g., a spoon, a hammer, and an eyeglass). Two candidate mechanisms are proposed. The prior-free mechanism posits that the new word is formed based only on the exemplars, while the prior-based Bayesian inference mechanism posits that structured priors construct the hypothesis space, and that learning a word concept engages updating this hypothesis space given the exemplars via Bayesian inference. **b** Participants ($N = 20$) underwent two fMRI experiments sequentially. In Experiment 1 (oddball one-back similarity judgment task, independent of the learning task), participants viewed each object and judged whether the object following a red fixation cross was similar to the previous object (1-s stimulus, 1–3 s fixation). In Experiment 2 (word concept learning task), in each trial, participants learned a new word concept from a few exemplars (2-s presentation, 0.5–1.5 s fixation) and then judged whether the new word concept could also be generalized to a probe object (2-s judgment phase). The key feature of this two-phase design is that neural representations of objects from Experiment 1 (independent of learning) were used to predict both the neural representations of new word concepts and the generalization behavior in Experiment 2. **c** Neural priors capture more complex information than behavior-judgment-based priors. Behavior-semantic (red outline, color-matched to the bar plot) and behavior-shape-based (blue outline) representational dissimilarity matrices (RDMs) were correlated

with neural RDMs in different brain regions. Partial correlations (controlling for competing RDMs) reveal distinct regional profiles: the higher-order ventral visual cortex (ventral occipitotemporal cortex, VOTC) correlated overall with the semantic RDM; bilateral fusiform (within the VOTC) correlated with both the semantic and shape-based RDMs; fusiform face area (FFA, within the VOTC) correlated selectively with the face-shape-based RDM; lateral occipitoparietal cortex (LOPC) correlated with the shape-based RDM. Regions were defined using the Harvard-Oxford Atlas, except that the FFA was defined by contrasting face pictures against animal/artifact pictures in Experiment 1 (120 top voxels around the peak voxel [40, -40, -22]; voxel-wise $T_s > 4.870$). Partial correlation analyses were one-tailed tests. No adjustment was made for multiple comparisons. Source data are provided as a Source Data file. *** $P < 0.001$; * $P < 0.05$. Silhouette images in (a, b) are made by yutl665 (www.flaticon.com/authors/yutl665), Smashicons www.flaticon.com/authors/smashicons, Freepik (www.flaticon.com/authors/freepik), LAFS, Pixel perfect (www.flaticon.com/authors/pixel-perfect), Maan Icons (www.flaticon.com/authors/maan-icons), Jesus Chavarria (www.flaticon.com/authors/jesus-chavarria), and Those Icons (www.flaticon.com/authors/those-icons) from www.flaticon.com under the Flaticon Free License. The novel-shape image is created by the authors and is available in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

hypothesis space of possible word meanings^{1,24,25}. Under this framework, learning involves evaluating each candidate hypothesis in the space, based on its initial plausibility (prior) and its ability to explain the observed exemplars (likelihood). Generalization to new instances is then governed by integrating these possible hypotheses and weighting each according to the learner’s posterior belief after observing the exemplars. Despite its simplicity, this probabilistic account—operating over a putative hypothesis space—captures key behavioral signatures of human word concept learning, including sensitivity to semantic structure and exemplar similarity, graded generalization after a single exemplar, and category-specific generalization with multiple exemplars. This provides initial yet important evidence for the role of existing object knowledge in word-concept learning.

Central to these Bayesian accounts is the assumption that the internal hypothesis space—whose properties critically shape how new word concepts are learned and generalized—is structured by learners’ existing conceptual knowledge. Previous work has estimated this space behaviorally, approximating its organization using relationships derived from participants’ similarity judgments of relevant objects (e.g., rating object similarity on a 7-point scale). While such one-dimensional behavioral judgments can reflect aspects of prior knowledge about object relationships, they inevitably yield a flattened projection of the rich and high-dimensional organization of semantic knowledge (see discussions in Binder et al.²⁶; Hebart et al.²⁷). Indeed, decades of neuroimaging and neuropsychological research on the neural basis of semantic memory have converged to reveal a distributed representation spanning the temporal, parietal, and frontal cortices and the underlying connections, supporting high-dimensional semantic knowledge representations (e.g., Fang et al.²⁸; Fernandino et al.²⁹; Binder et al.³⁰). For object knowledge, for instance, the higher-order ventral visual cortex (ventral occipitotemporal cortex, VOTC) has been shown to contain complex representations of various types of object attributes (e.g., shape and color), proposed as semantic grounding in perceptual/experiential systems^{26,29,31–34}, and to respect an evolutionarily salient domain principle (e.g., the animate-inanimate principle reflected by the organized representations of animals, human faces and artifacts in the human brain^{35–38}). Therefore, it is possible that the rich, structured, and distributed neural representations of object knowledge serve as a neurobiological basis for prior beliefs central to Bayesian learning. However, which brain regions encode prior knowledge in word concept learning, and by what neural mechanisms this information guides word concept learning and generalization, remain elusive.

To bridge the gap between the neural foundation of prior knowledge and the theoretical framework of word concept learning, we propose a Neural Bayesian Model (NBM) that addresses how cortical knowledge structures guide word concept learning. NBM is a neurobiologically grounded model of word concept learning where the hypothesis space for Bayesian inference is constructed directly from neural representations of long-term object knowledge. This approach provides an objective, high-resolution specification of the priors underlying Bayesian learning. Critically, it enables, within a unified experimental and computational framework, a systematic investigation of how word concept learning is shaped by different priors encoded in brain regions previously implicated in distinct aspects of conceptual processing. Specifically, the NBM allows us to examine how priors encoded in the VOTC (implicated in object knowledge representation) and the hippocampus, VMPFC, and DMPFC (implicated in learning novel associations and concepts) differentially support word concept learning and generalization, potentially reflecting their complementary neurocomputational roles.

We applied the NBM to address five key questions (see Table 1 for a summary of corresponding analytical approaches). First, do neural representations in the VOTC, hippocampus, VMPFC, and DMPFC provide structured priors in support of learning a new word (e.g., “Leca”) via exemplars (e.g., a spoon, an eyeglass, and a hammer)? Second, are the structured neural priors beyond specific exemplars necessary for forming the neural representations of the learned words in these regions? Third, to what extent does learning with familiar objects (with robust and rich priors) differ from learning with novel objects (with weak priors)? Fourth, is word learning differentially supported by the VOTC and the hippocampus/MPFC? Fifth, do large language models (LLMs)—a system endowed with extremely rich semantic priors without explicitly incorporating the Bayesian learning mechanism—naturally emerge human-like word concept learning behavior? Together, these questions position the proposed NBM model as a formal framework for characterizing the distinct contributions of prior-based mechanisms in word concept learning.

Results

The results are structured into the following sections for constructing and evaluating the NBM: (1) Developing the NBM by neural prior measurement (Fig. 1) and prediction computation (Fig. 2); (2) Evaluating the effectiveness of the NBM, for regions of interest (VOTC and hippocampus/VMPFC/DMPFC) separately, in predicting the neural representations and behavior (for VOTC, Figs. 3 and 4; for

Table 1 | Summary of our study’s five key questions and corresponding analytical approaches

Key question	Analytical approach
Q1: Do brain regions involved in knowledge and learning provide structured neural priors for concept learning?	Predicting word concept neural representations and generalization behavior based on region-specific NBM.
Q2: Are structured neural priors necessary for concept learning?	Comparing NBM to models with no prior, permuted prior, and behavior-prior.
Q3: Does prior richness modulate learning?	Comparing NBM’s prediction for familiar-object learning with that for novel-shape learning at both behavioral and neural levels.
Q4: Do the VOTC and hippocampus/MPFC support learning differently?	Testing patterns between brain region and concept type (familiar objects/novel shapes).
Q5: Do multimodal LLMs with rich priors show human-like learning patterns?	Comparing the human-learning behavior prediction performance of LLMs and the Bayesian learning model.

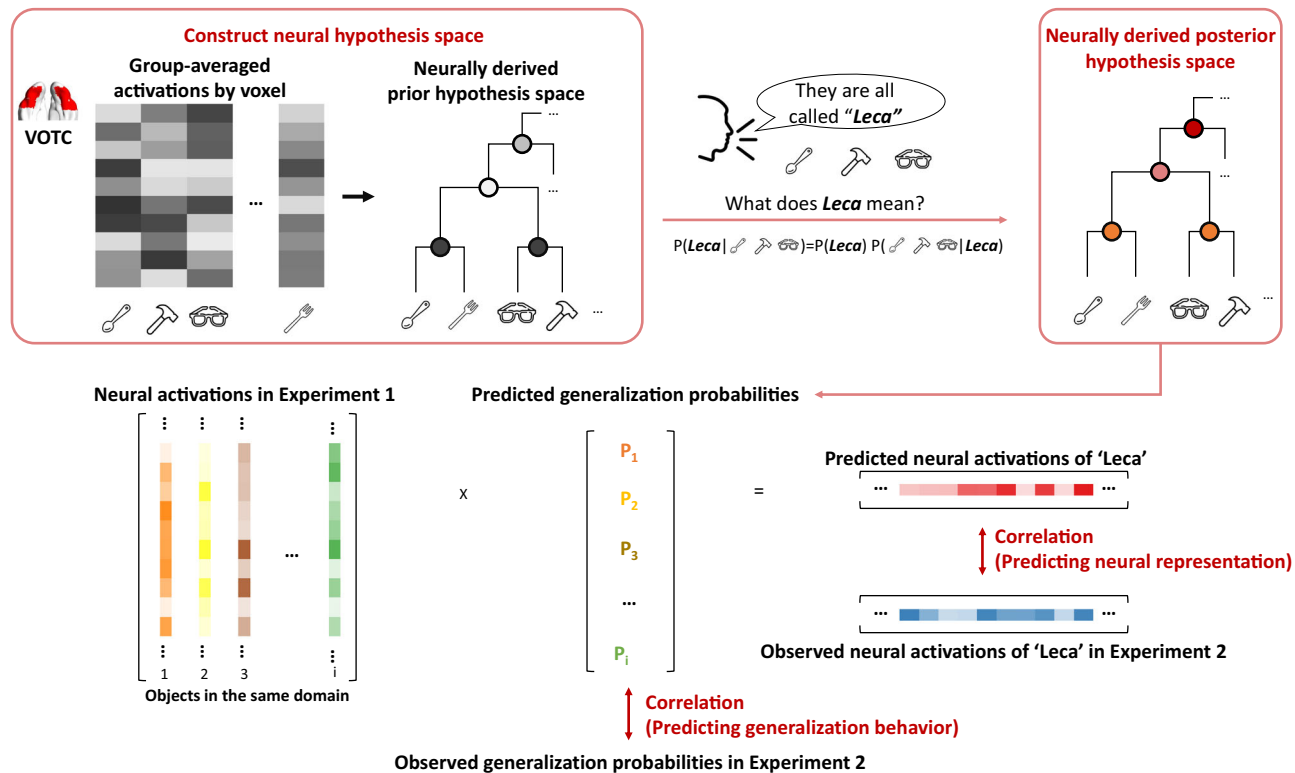


Fig. 2 | Flowchart of Neural Bayesian Model (NBM) construction and prediction (illustrated for “Leca”). The key procedure is to construct the NBM based on the group-averaged activation patterns in Experiment 1, then to predict the word concept neural representation and generalization behavior in Experiment 2. To construct the NBM, for a given brain region, hierarchical analysis is applied to the group-mean activation patterns (t-values) of the 58 object stimuli. The resulting dendrogram defines the neurally derived prior hypothesis space (see Fig. S5 for the full dendrograms based on neural priors in the ventral occipitotemporal cortex (VOTC) and on behavioral priors). Each node in the dendrogram represents a possible hypothesis. Given the exemplars presented during learning, Bayesian inference updates the prior hypothesis space to a posterior distribution. From this posterior, the probability that a probe object belongs to “Leca” (i.e., the

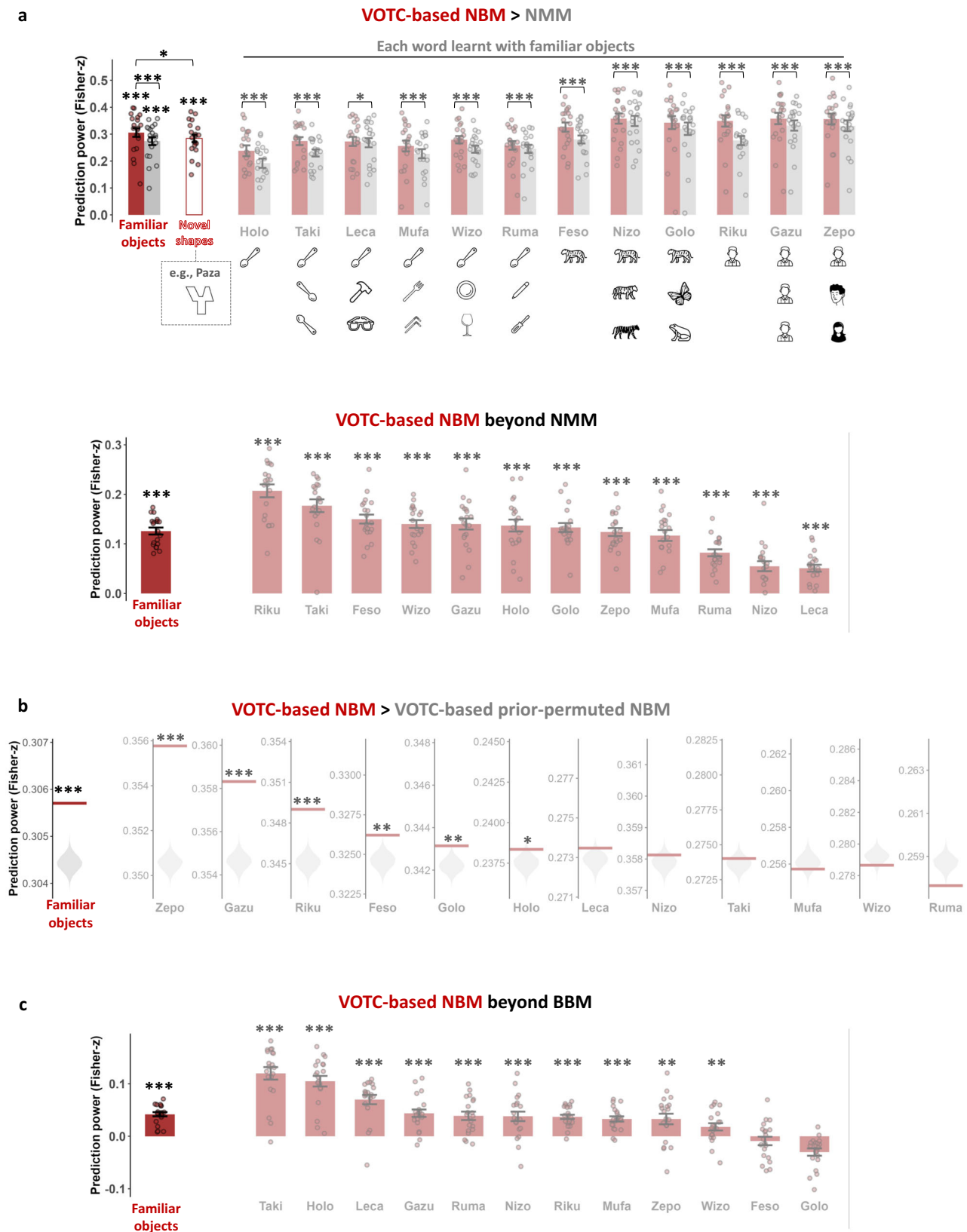
generalization probability) is computed. The neural pattern of “Leca” is predicted as the average of the neural activation patterns of objects in the same domain obtained in Experiment 1, weighted by the corresponding generalization probabilities. The correlations between the predicted and observed neural activation patterns (and generalization probabilities) of new words are calculated to evaluate the predictive power of the NBM, with higher correlations indicating better model performance. Silhouette images are made by yut1665 (www.flaticon.com/authors/yut1665), Smashicons (www.flaticon.com/authors/smashicons), Freepik (www.flaticon.com/authors/freepik), Pixel perfect (www.flaticon.com/authors/pixel-perfect), and Maan Icons (www.flaticon.com/authors/maan-icons) from www.flaticon.com under the Flaticon Free License.

hippocampus/VMPFC/DMPFC, Fig. 5) of new word learning for familiar objects with rich priors, against multiple control models (prior-free model; permutation model; behavioral prior model) and conditions without structured neural prior considerations (weak prior stimuli–novel shapes); (3) Comparing the Bayesian learning model with LLMs in fitting human word concept learning behavior (Fig. 6).

Overview of the experimental procedures

To obtain neural representational priors, we first measured the brain activity of a set of objects covering the object domains that have been

well studied ($N = 58$) (fMRI Experiment 1; see Fig. S1 for all the stimuli). These object-evoked neural responses reflect aspects of knowledge stored in participants’ long-term (semantic) memory, providing the basis for constructing a neural hypothesis space for the NBM. A subset of these objects was then used as exemplars in a subsequent word concept learning experiment conducted inside the scanner (fMRI Experiment 2). See Fig. 1b for an overview of the experimental paradigms. In each trial, participants were first introduced to a novel word (e.g., “Leca”) together with several exemplars (e.g., a spoon, an eyeglass, and a hammer), and were then asked to judge whether a probe object (e.g., an axe) could also



be described by the new word. These generalization judgments had no correct or incorrect answer, allowing us to measure subjective interpretations as the observed generalization probabilities.

The key aspect of this two-phase design is that we used the neural representations of each object obtained from Experiment 1, which was

independent of the learning task, to construct the priors of the NBM and then to predict both the neural representations of new word concepts and the generalization behavior obtained from Experiment 2 (i.e., the learning task). See Fig. 2 for a schematic of the analysis pipeline.

Fig. 3 | The VOTC-based NBM predicts neural representations of word concepts in the VOTC. In each panel, the leftmost bar/violin (labeled “Familiar objects” in red) shows the combined results across all familiar-object word concepts. The remaining bars/violins (faded and labeled in gray) show results for individual familiar-object word concepts. Each point in the bar plot represents the data from one participant (same in all panels). **a** The VOTC-based NBM outperforms the neural mean model (NMM, based on the prior-free mechanism; gray bars) and has unique predictive power (after controlling for the NMM’s prediction) for familiar-object word concepts. For novel-shape learning, the VOTC-based NBM also yields significant predictions (white bar with red outline) but performs worse than for familiar-object learning, indicating that the NBM is especially suited to learning with rich priors. Exemplars for each word are shown below the corresponding bar. Analyses were conducted using one-sample *t* tests (one-tailed for single models, and two-tailed for model and condition comparisons). **b** The VOTC-based NBM outperforms the prior-permuted NBM. Violin plots show the null distributions; red lines indicate the observed group-mean predictive power of the NBM. Analyses were nonparametric tests based on permutation (one-tailed). **c** The VOTC-based

NBM has unique predictive power beyond the BBM. Analyses were conducted using one-sample *t* tests (one-tailed). No adjustment was made for multiple comparisons. VOTC ventral occipitotemporal cortex, NBM neural Bayesian model, NMM neural mean model, BBM behavioral Bayesian model. Source data are provided as a Source Data file. ****P* < 0.001; ***P* < 0.01, **P* < 0.05. Silhouette images in (a) are made by Smashicons (www.flaticon.com/authors/smashicons), Freepik (www.flaticon.com/authors/freepik), Pixel perfect (www.flaticon.com/authors/pixel-perfect), Maan Icons (www.flaticon.com/authors/maan-icons), Good Ware (www.flaticon.com/authors/good-ware), Amethyst prime (www.flaticon.com/authors/amethyst-prime), Nadiinko (www.flaticon.com/authors/nadiinko), monkik (www.flaticon.com/authors/monkik), khld939 (www.flaticon.com/authors/khld939), Irfansanto20 (www.flaticon.com/authors/irfansanto20), Jesus Chavarria (www.flaticon.com/authors/jesus-chavarria), and AB Design (www.flaticon.com/authors/ab-design) from www.flaticon.com under the Flaticon Free License. The novel-shape image is created by the authors and is available in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

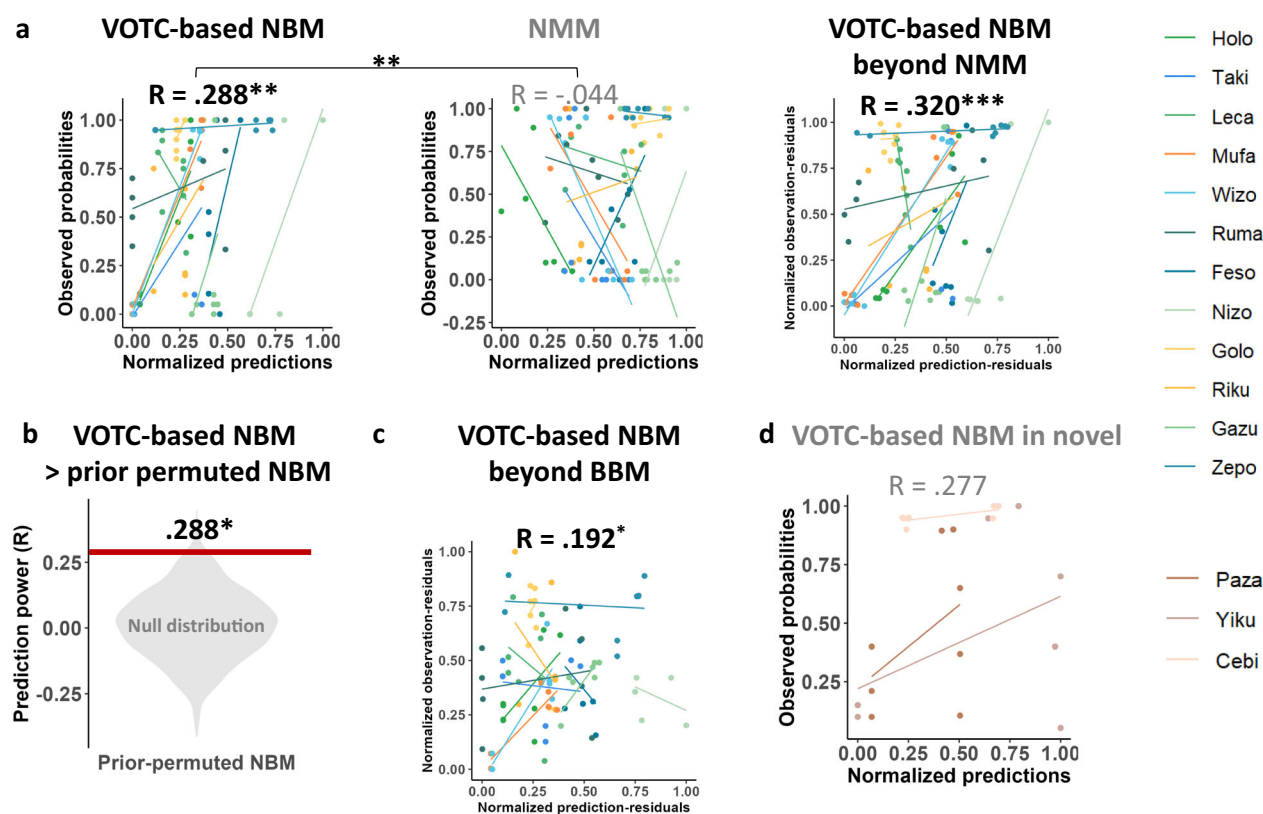


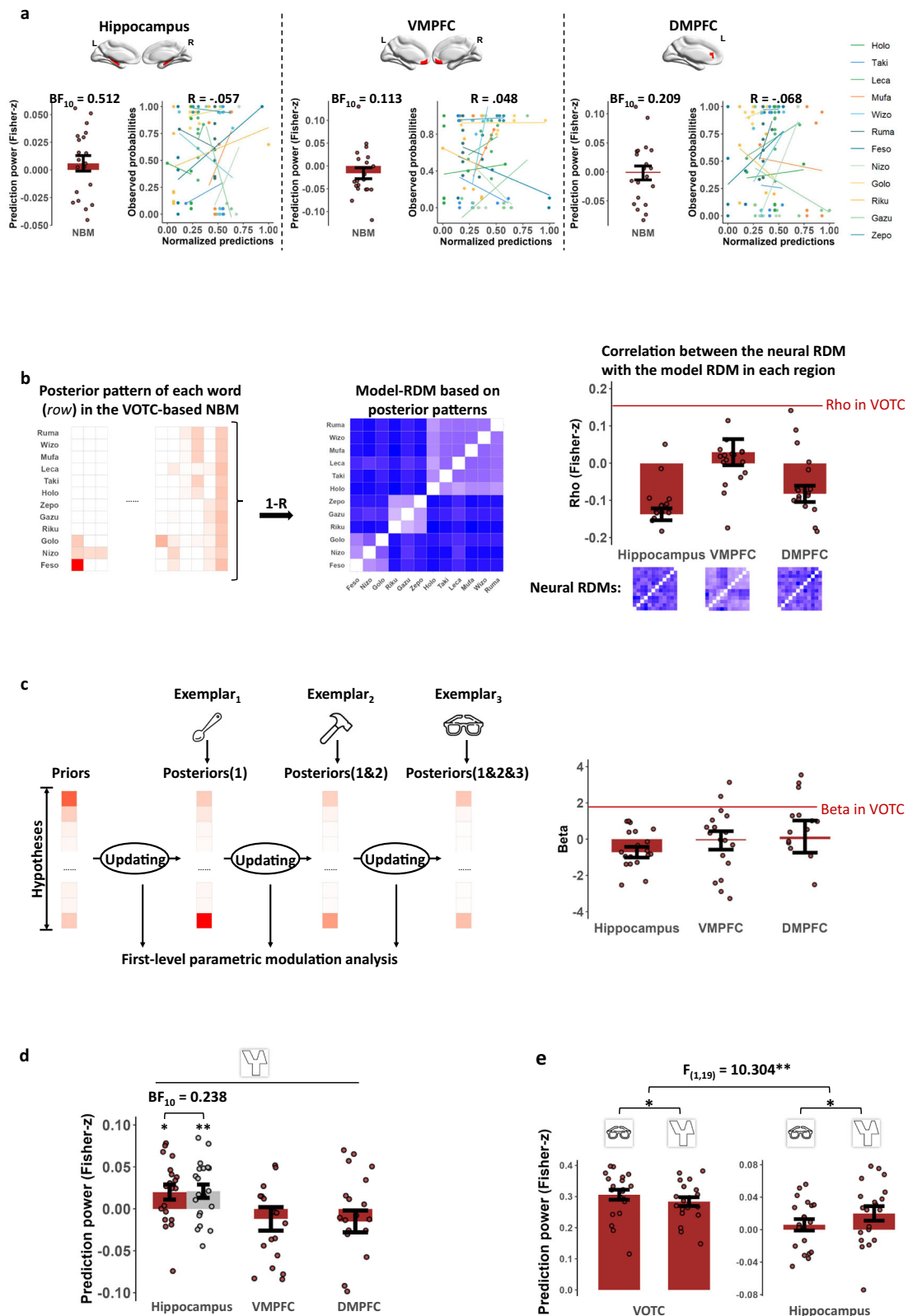
Fig. 4 | The VOTC-based NBM predicts the generalization behavior for familiar-object learning. Each point in the scatter plots represents the data from one unique trial (a unique word–probe combination), with colors distinguishing trials from different words. The *R*-value above each scatterplot is the correlation across all trials; nonsignificant correlations are displayed in gray. Predicted values from each model were normalized to the range 0–1 across trials. **a** The VOTC-based NBM outperforms the NMM and has unique predictive power beyond the NMM for familiar-object word concepts. Correlation analyses were one-tailed tests. Model comparisons were

conducted using Filon’s *Z*-test (two-tailed). **b** The VOTC-based NBM outperforms the prior-permuted NBM. This analysis was a nonparametric test based on permutation (one-tailed). **c** The VOTC-based NBM has unique predictive power beyond the BBM. This correlation analysis was a one-tailed test. **d** The VOTC-based NBM did not significantly predict generalization behavior for novel-shape learning. No adjustment was made for multiple comparisons. VOTC ventral occipitotemporal cortex, NBM neural Bayesian model, NMM neural mean model, BBM behavioral Bayesian model. Source data are provided as a Source Data file. ****P* < 0.001, ***P* < 0.01, **P* < 0.05.

Behavior results

Participants showed high intersubject consistencies (ISCs) in behavioral responses in both experiments (Experiment 1: oddball object similarity judgment task, Fisher-transformed ISC = 0.517, *P* < 0.001; Experiment 2: word concept learning task, Fisher-transformed ISC = 0.483, *P* < 0.001; see Section A1 in the Supplementary Information for details). Critically, in the word concept learning task (Experiment 2), the generalization behavior replicated the classical result pattern¹, with sensitivity to the

number and similarity of the referent exemplars (Fig. S2a). When learning with a single exemplar, participants showed graded generalization to new objects that had high, medium, and low similarity to the learned exemplar; when learning with three exemplars, participants showed generalization sharpened into a more all-or-none pattern, depending on the similarity among the three exemplars. We also replicated the predictive power of the behavioral Bayesian model (BBM) constructed using behavioral semantic judgments for the



generalization behavior (Fig. S2b; Pearson $R = 0.907$, 95% confidence interval = $[0.871, 1]$, $BF_{10} = 5.779 \times 10^{32}$, $T_{94} = 20.847$, $P_{one-tailed} = 2.54 \times 10^{-37}$; Section A2 in the Supplementary Information).

NBM

The NBM formalizes word concept learning as a probabilistic inference over a structured neural hypothesis space (see Fig. 2 for a flowchart of

the NBM construction). According to the model, learning a word such as “Leca” involves evaluating a range of candidate hypotheses about its potential referents. Each hypothesis corresponds to a potential category in semantic memory (e.g., all spoons, all tableware, or all artifacts for “Leca”). These possibilities are structured hierarchically, forming an internal hypothesis space shaped by prior neural representations of object relationships.

Fig. 5 | Different patterns of the VOTC and hippocampus in word concept learning driven by neural priors. In bar plots, each point represents the data from one participant. In scatter plots, each point represents the data from one trial (a unique word-probe combination), with colors distinguishing trials from different words. **a** Using neural priors from each region itself, the NBM did not significantly predict new word neural representations (bar plots) or generalization behavior (scatter plots) for familiar-object learning in the hippocampus, VMPFC, or DMPFC. The one-sample t tests and correlation analyses were one-tailed. **b** Using VOTC-derived priors, the NBM did not significantly predict new word neural representation similarities in the hippocampus, VMPFC, or DMPFC for familiar-object learning. The model RDM based on the VOTC representations was not significantly correlated with neural RDMs in the hippocampus, VMPFC, or DMPFC. The red line in the bar plot indicates the correlation with neural RDMs in the VOTC. Analyses were conducted using one-sample t tests (one-tailed). **c** Using VOTC priors, the NBM did not significantly track word-concept representation updating in the hippocampus, VMPFC, or DMPFC during familiar-object learning. Updating was quantified with the VOTC-based NBM and regressed against activation in the hippocampus, VMPFC, and DMPFC via parametric modulation; no significant relationships were found. The red line in the bar plot indicates the modulation

results in the VOTC. Analyses were conducted using one-sample t tests (one-tailed). **d** The hippocampus supports novel-shape concept learning. For novel-shape learning, the NBM predicted hippocampal neural representations and did not differ from the NMM (gray bar). Analyses were conducted using one-sample t tests (one-tailed for single models, and two-tailed for model comparison). **e** Two-way ANOVA (two-tailed) showed significant interaction between region and word type: the NBM shows better predictive power for familiar-object than novel-shape learning in the VOTC, but better predictive power for novel-shape than familiar-object learning in the hippocampus. Simple effects were tested with one-sample t tests (two-tailed). No adjustment was made for multiple comparisons. VOTC ventral occipito-temporal cortex, VMPFC ventromedial prefrontal cortex, DMPFC dorsomedial prefrontal cortex, NBM neural Bayesian model, NMM neural mean model. Source data are provided as a Source Data file. ** $P < 0.01$, * $P < 0.05$. Silhouette images in (c–e) are made by Smashicons (www.flaticon.com/authors/smashicons), Freepik (www.flaticon.com/authors/freepik), and Pixel perfect (www.flaticon.com/authors/pixel-perfect) from www.flaticon.com under the Flaticon Free License. The novel-shape image is created by the authors and is available in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

Before observing any learning exemplars, the learner assigns each hypothesis a prior probability based on how likely that category is to be named. In line with previous work¹, we assume that these priors are derived from the structure of the hypothesis space; more distinctive categories are considered more name-worthy. Formally, the prior for each hypothesis is proportional to the height difference between the hypothesis and its parent in the tree-structured hypothesis space. For instance, if “tableware” forms a well-separated cluster in the tree, it is assigned a larger prior than a less distinctive category, such as “kitchen items.” This reflects the intuition that some categories are more likely to be lexicalized than others.

When learners observe labeled exemplars (e.g., a spoon, an eye-glass, and a hammer) for a new word (e.g., “Leca”), they update their beliefs using Bayes’ theorem¹. The likelihood captures how well each hypothesis explains the observed exemplars, favoring smaller categories that tightly encompass the exemplars. Formally, the model assumes that the likelihood is inversely proportional to the size of the hypothesis (i.e., the cardinality). The posterior then combines the prior and likelihood to yield updated beliefs about the probable meaning.

Generalization involves integrating all plausible hypotheses¹. When deciding whether a new object (e.g., an axe) could also be a “Leca”, the learner does not choose a single hypothesis with the highest posterior. Instead, they sum the probabilities of all hypotheses containing both the new object (e.g., the axe) and the observed exemplars within the internal hypothesis space.

The neural representation of a new word was then predicted by averaging the prior neural representations of corresponding objects of the same domain (obtained in Experiment 1), weighted by their generalization probabilities (“Methods”).

The key idea of the NBM is to construct a prior hypothesis space directly from neural representations of long-term object knowledge and to use this space to guide Bayesian word concept learning. We first obtained multivoxel response patterns for 58 objects across participants in Experiment 1, which was independent of the word concept learning task (Experiment 2). Within each region of interest (ROI), we averaged the neural response elicited by each object across all participants. The object-specific response patterns within the ROI were then used to construct a representational dissimilarity matrix (RDM), capturing the similarity structure among the objects as represented in the ROI. Subsequently, we applied a standard hierarchical clustering algorithm to the neural RDM in the ROI to derive a dendrogram structure that functioned as a prior hypothesis space for Bayesian word concept learning. Each node in this tree corresponds to a possible hypothesis about a word concept (e.g., the spoon and fork are

“tableware”), with more specific hypotheses nested under broader ones.

To validate the approach used to derive the neural hypothesis space, we conducted a series of checks. As a sanity check, we first examined the univariate activation maps in Experiment 1. As expected, a contrast analysis of object categories revealed robust category-selective activations, aligning well with the literature (Fig. S3 and Table S1): face stimuli elicited stronger responses in regions including the right fusiform cortex (encompassing the fusiform face area), animal pictures activated regions extending from the bilateral posterior fusiform cortex to the lateral occipital cortex, and artifact pictures elicited activity in the bilateral lateral superior occipital cortex, among other regions^{35–38}.

To assess the representational content captured by the neural RDM, and to test whether it encodes richer information than behavioral judgments, we compared brain responses with behavioral similarity judgments along two important object properties—shape and semantic meaning. We examined the representational patterns in several visual cortex regions by taking advantage of the fact that long-term object knowledge is encoded in a distributed manner across the visual cortex. Participants judged the pairwise distance (i.e., dissimilarity) of objects based on semantic and shape-based dimensions, yielding behavioral RDMs for each (Fig. 1c).

The two object-property RDMs were significantly correlated ($Rho = 0.668$). Using representational similarity analysis (RSA), we found that neural responses in different visual cortex regions represented these two types of object properties to different degrees: the VOTC, defined by the Harvard-Oxford Atlas (see “Methods”), overall showed a unique effect of the behavioral-semantic RDM (partial $Rho = 0.286$ with the shape RDM and the low-level image (i.e., pixel-based) RDM controlled for, 95% confidence interval = [0.237, 1], $T_{986} = 9.369$, $P_{one-tailed} = 2.43 \times 10^{-20}$); examining subsections of the VOTC, the neural RDMs of bilateral fusiform (defined by the temporal occipital fusiform cortex in the Harvard-Oxford Atlas) showed unique correlations with both the behavioral-semantic (partial $Rho = 0.231$, 95% confidence interval = [0.181, 1], $T_{986} = 7.460$, $P_{one-tailed} = 9.47 \times 10^{-14}$) and -shape RDMs (partial $Rho = 0.213$, 95% confidence interval = [0.162, 1], $T_{986} = 6.839$, $P_{one-tailed} = 6.99 \times 10^{-12}$); the fusiform face area, defined by contrasting face pictures with animal and artifact pictures (120 top voxels around the peak voxel [40, -40, -22]; voxel-wise $T_s > 4.870$), showed a unique correlation with the behavioral-shape RDM of face pictures (partial $Rho = 0.203$, 95% confidence interval = [0.013, 1], $T_{74} = 1.779$, $P_{one-tailed} = 0.040$) but not for artifact pictures (partial $Rho = -0.085$, 95% confidence interval = [-0.210, 1], $T_{167} = -1.100$, $P_{one-tailed} = 0.864$).

a Large language models to be tested: GPT-4o and Qwen2.5-VL

Prompt procedures:

system prompt: 你只能回答“是”或“否”，绝对不能回答其他内容。['You can only answer "Yes" or "No", and absolutely not anything else.']

假如你有一位说另一种语言的朋友，这位朋友看到这一张图中的物体，把它称作'Leca' (无需回复)
['Supposing that your foreign friend who speaks another language sees this object in this picture and calls it 'Leca' (no need to respond)']



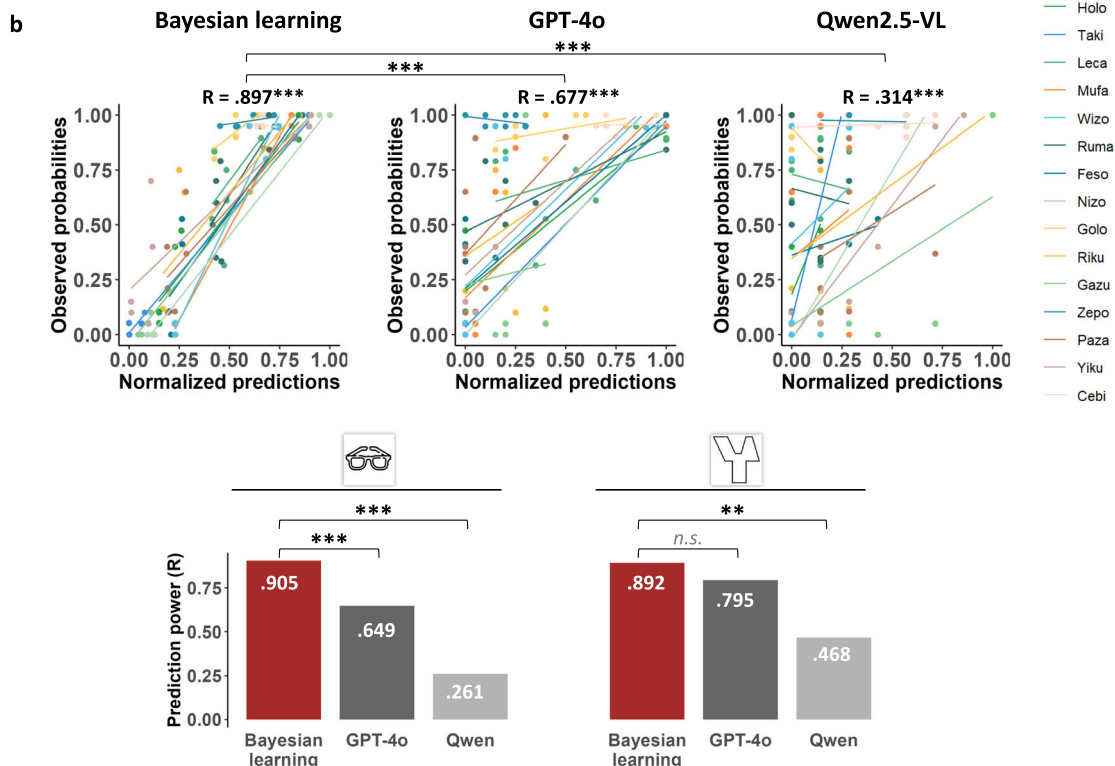
假如你有一位说另一种语言的朋友，这位朋友看到这一张图中的物体，把它称作'Leca' (无需回复)
['Supposing that your foreign friend who speaks another language sees this object in this picture and calls it 'Leca' (no need to respond)']



假如你有一位说另一种语言的朋友，这位朋友看到这一张图中的物体，把它称作'Leca' (无需回复)
['Supposing that your foreign friend who speaks another language sees this object in this picture and calls it 'Leca' (no need to respond)']



请问这位朋友看到这张图中的物体时，是否也把它称作'Leca'，请给出是或否的判断
['When this friend sees this object in this picture, would they also call it a 'Leca'? Please answer with 'yes' or 'no'.']



By contrast, the bilateral lateral occipitoparietal cortex (defined by the superior lateral occipital cortex in the Harvard-Oxford Atlas) showed a unique correlation with behavioral-shape RDMs (partial $Rho = 0.292$, 95% confidence interval = [0.244, 1], $T_{986} = 9.608$, $P_{one-tailed} = 3.07 \times 10^{-21}$). Although the RSA results with these specific behavioral RDMs indicate different representational contents across regions, it is important to note that these behavioral models captured only a

small fraction of the neural RDM variance. Neural RDMs may contain, beyond neural activity-measuring noise, prior representational contents about objects beyond the specific articulated dimensions that the behavioral judgments probe, hence emphasizing the value of neural priors. The neural RDMs were also computed for the hippocampus, VMPFC, and DMPFC (Fig. S4), and no significant correlation was observed with either behavioral-semantic or behavioral-shape RDMs.

Fig. 6 | Large language models (LLMs) are less human-like than the Bayesian learning model during word concept learning. **a** Two multimodal LLMs, GPT-4o and Qwen2.5-VL, were prompted to perform the same word-learning task as human participants. For a given trial, exemplars were presented in three separate messages, mirroring the setting in human experiments. All prompts were in Chinese to maintain consistency with the human-learning task. Each trial was repeated 20 times in an independent chat session, with temperature = 1.0 and top- $P = 1.0$ to sample from the models' original output distributions. Generalization probability for a given trial was defined as the proportion of “yes” responses across the 20 iterations. **b** Although both LLMs significantly predict generalization behavior, the Bayesian learning model outperforms them. This advantage persisted for familiar-object and novel-shape learning, except that the Bayesian model did not significantly differ from GPT-4o for novel-shape learning. The Bayesian learning model's predictions were obtained by regressing the observed generalization

behavior onto the predictions of both the VOTC-based NBM and the BBM using a general linear model, via a leave-one-trial-out approach. All model predictions were normalized to the range 0–1 across trials. Each point in the scatterplot represents the data from one trial (a unique word–probe combination), with colors distinguishing trials from different words. Correlation analyses were one-tailed. Model comparisons were conducted using Filon's Z-test (two-tailed). No adjustment was made for multiple comparisons. Source data are provided as a Source Data file. *** $P < 0.001$; ** $P < 0.01$. Silhouette images are made by Smashicons (www.flaticon.com/authors/smashicons), Freepik (www.flaticon.com/authors/freepik), Pixel perfect (www.flaticon.com/authors/pixel-perfect), and Those Icons (www.flaticon.com/authors/those-icons) from www.flaticon.com under the Flaticon Free License. The novel-shape image is created by the authors and is available in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

What does the neurally derived hypothesis space look like, and how does it compare to the hypothesis space built from behavioral judgments? We compared the dendrogram computed based on the RDM in the VOTC with that derived from behavioral similarity judgments, following the same hierarchical clustering procedure. As shown in Fig. S5, the two dendrograms reveal broadly similar hierarchical structures. For instance, both trees clearly separate major object domains (i.e., faces, animals, and artifacts). However, they also exhibit notable fine-grained differences. For example, while the spoon (object 19) and pen (object 20) are clustered closely together in the VOTC-based dendrogram, they are placed in distinct sub-branches in the behavioral hypothesis space. These differences highlight that the priors reflected by brain activity may capture subtle latent information in object representations that diverges from explicit human similarity judgments, highlighting their value in revealing internal knowledge representations that are likely to inform word concept learning.

Word concept learning with rich priors in the VOTC

We evaluated the NBM's effectiveness in predicting neural representations and behavioral patterns via: (1) the correlation between predicted and observed neural representations of words, and (2) the correlation between predicted generalization probabilities for new probe objects and the observed behavioral probabilities. Note that to exclude the possibility that the NBM's potential prediction success is simply due to correlations between brain responses to the same pictures in Experiment 1 and 2, and to ensure that the NBM's performance indeed benefits from the neural prior-based Bayesian inference, the performance of the NBM was further compared to a series of control models (Figs. 3 and 4; see below for details of the rationale and construction of each control model).

In the first analysis (i.e., predicting neural representations), the predictions of VOTC-based NBM for word concepts were highly significant (Pearson $R_{\text{Fisher-Z}} = 0.306$, lower boundary of 95% confidence interval = 0.277, Cohen's $d = 4.143$, $BF_{10} = 9.630 \times 10^{10}$, $T_{19} = 18.529$, $P_{\text{one-tailed}} = 6.36 \times 10^{-14}$; Fig. 3a; for the statistical results of this and the following analyses, see Table S2). Thus, a Bayesian model based on abstract structural properties of neural representations in the VOTC significantly predicted the neural encoding of meaning for a newly learned word.

We chose items from three major domains whose representational distinctions are consistently observed in ventral visual and semantic systems (animals, human faces, artifacts), i.e., with qualitatively different neural priors^{35–38}. Unless noted otherwise, results are reported collapsed across domains for presentation simplicity in the main text. Importantly, domain-specific analyses yielded the same qualitative pattern (for NBM's prediction in each domain, $T_{S19} > 17$, $Ps < 0.001$; Table S3).

To evaluate the contribution of structured neural priors, we compared the NBM with a set of control models and a control learning condition: (1) a model derived from a prior-free mechanism, (2) models

constructed based on randomly permuted neural priors, (3) a model constructed based on behavioral-judgment priors, and (4) a learning condition with weak priors.

We compared the NBM with a key control model that does not incorporate structured priors but still uses neural signals. This model predicts the learned word concept representation by simply averaging the neural activation patterns of its associated exemplars without considering their distribution in a broader hypothesis space, hence called the Neural Mean Model (NMM). For example, for the word “Leca”, which was associated with the picture exemplars spoon, eye-glass, and hammer in Experiment 2, the NMM predicted the neural representation of this new word as the mean of the neural patterns of these three objects independently measured in Experiment 1. If cross-experiment prediction were driven solely by the neural activity to the same pictures, rather than incorporating neural prior structures, the NMM would be expected to outperform or at least match the NBM. Indeed, the prediction of the NMM was significant (Fig. 3a, top; Pearson $R_{\text{Fisher-Z}}$ between predicted and observed neural activity pattern in the VOTC = 0.274, lower boundary of 95% confidence interval = 0.248, Cohen's $d = 4.091$, $BF_{10} = 7.753 \times 10^{10}$, $T_{19} = 18.294$, $P_{\text{one-tailed}} = 8.00 \times 10^{-14}$).

Critically, the predictive power of the NBM was significantly greater than that of the NMM (Fig. 3a, top; Pearson $R_{\text{Fisher-Z}} = 0.306$ vs. 0.274, 95% confidence interval = [0.027, 0.035], Cohen's $d = 3.724$, $BF_{10} = 7.934 \times 10^9$, $T_{19} = 16.653$, $P_{\text{two-tailed}} = 8.63 \times 10^{-13}$). In addition, after controlling for the predicted neural pattern of the NMM via partial correlation, the NBM still had unique predictive power (Fig. 3a, bottom, Pearson $R_{\text{Fisher-Z}} = 0.126$, lower boundary of 95% confidence interval = 0.115, Cohen's $d = 4.327$, $BF_{10} = 1.011 \times 10^{11}$, $T_{19} = 19.353$, $P_{\text{one-tailed}} = 2.89 \times 10^{-14}$). These results indicate that the NBM's predictions extend beyond exemplar-specific visual information contained in the measured activations, and that incorporating broader, structured neural priors indeed contributes to the NBM's predictive power.

To test whether the advantage of the NBM was indeed driven by incorporating specific neural priors, we compared the performance of the NBM to a prior-permuted control. In this control, object identities were randomly shuffled, disrupting the structure of the neural priors, before constructing the hypothesis dendrogram used for Bayesian learning. A null distribution was generated by running 100 permutation iterations and 10,000 group-level bootstrap samples.

The predictive power of the NBM constructed from the actual neural priors significantly outperformed the prior-permuted control models (Fig. 3b; observed Pearson $R_{\text{Fisher-Z}} = 0.306$, mean of the null distribution = 0.304, $P_{\text{one-tailed}} = 1.88 \times 10^{-14}$). The Standard Effect Size (SES) was calculated as the difference between the observed value and the mean value of the null distribution, divided by the standard deviation of the null distribution, and was 7.569. That is, specific, structured neural priors are crucial for the NBM's predictive power.

Neural priors have the benefit of respecting the distributed, high-dimensional representation properties of priors, as opposed to

behavior-judgment-based priors, which are obtained through judgment along one dimension or a black-box composite (Fig. 1c). To test whether the NBM indeed has additional predictive power, we compared it with the BBM, which was constructed using the behavioral judgments of semantic distance of sample objects.

The results showed that after controlling for the predicted pattern from the BBM using partial correlation, the NBM had unique predictive power for the learned word neural representation in the VOTC (Fig. 3c; Pearson $R_{\text{Fisher-z}} = 0.042$, lower boundary of 95% confidence interval = 0.034, Cohen's $d = 2.246$, $BF_{10} = 4.851 \times 10^6$, $T_{19} = 10.043$, $P_{\text{one-tailed}} = 2.46 \times 10^{-9}$).

Another way to test the advantage of the NBM driven by the incorporation of neural priors is to compare learning with and without rich priors. We compared the predictive power of the NBM in learning with familiar objects (i.e., with rich, structured prior representations) with that in learning with novel shapes (i.e., with weak prior representations).

The results showed that although the predictive power of the NBM in learning with novel shapes was significant (Fig. 3a, top; Pearson $R_{\text{Fisher-z}} = 0.284$, lower boundary of 95% confidence interval = 0.261, Cohen's $d = 4.599$, $BF_{10} = 5.735 \times 10^{11}$, $T_{19} = 20.568$, $P_{\text{one-tailed}} = 9.57 \times 10^{-15}$), the NBM in learning with familiar objects outperformed the NBM in learning with novel shapes (Pearson $R_{\text{Fisher-z}} = 0.306$ vs. 0.284, 95% confidence interval = [0.005, 0.037], Cohen's $d = 0.627$, $BF_{10} = 4.606$, $T_{19} = 2.802$, $P_{\text{two-tailed}} = 0.011$). Thus, the NBM exhibits greater predictive power for learning with richer priors.

We conducted linear mixed-effects model analyses in parallel to control for the random effects of word concepts and participants. The results were consistent with those reported above (for all model comparisons, Cohen's $ds > 1$, $Ps < 0.001$; for comparison between familiar-object and novel-shape learning, $b = 0.021$, 95% confidence interval = [0.005, 0.038], standardized coefficient = 0.095, $P = 0.013$; see Section B4 in the Supplementary Information for details).

In the second analysis, we evaluated the model performance by computing the Pearson correlation between the NBM-predicted generalization probabilities and the actual behavioral responses (i.e., the percentage of participants who judged that the probe object belonged to the learned word). The results showed that the NBM significantly predicted generalization behavior (Fig. 4a, left; Pearson $R = 0.288$, 95% confidence interval = [0.125, 1], $BF_{10} = 21.466$, $T_{94} = 2.915$, $P_{\text{one-tailed}} = 0.002$). Similar to the neural analyses above, we compared the NBM's predictive power to that of the same control models and learning condition, revealing consistent results.

We considered the relative strength of the NBM, compared with neural models without prior incorporation (the NMM) and with random prior representations (prior-permuted control models). In the NMM, the generalization probability to a probe object is predicted as the Pearson correlation between the neural pattern of the probe object (obtained in Experiment 1) and the NMM-predicted neural pattern of the new word. Correlating these predicted values with the observed human behavior revealed that the NMM failed to yield effective predictions (Fig. 4a, middle; Pearson $R = -0.044$, 95% confidence interval = [-0.242, 1], $BF_{10} = 0.173$, $T_{94} = -0.429$, $P_{\text{one-tailed}} = 0.666$). A direct magnitude comparison showed the advantage of the NBM over the NMM ($z = 2.927$, $P_{\text{two-tailed}} = 0.003$). After controlling for the NMM's predicted values using partial correlation, the NBM still yielded significant predictions (Fig. 4a right; Pearson $R = 0.320$, 95% confidence interval = [0.159, 1], $BF_{10} = 54.923$, $T_{94} = 3.270$, $P_{\text{one-tailed}} = 7.52 \times 10^{-4}$). Examining the power of the NBM relative to prior-permuted control models (the same as above, but with 1000 permutation iterations and without group-level bootstrapping) again revealed the significant advantage of the NBM (Fig. 4b; mean of the null distribution = 0.014, $SES = 2.048$, $P_{\text{one-tailed}} = 0.020$). These results supported that incorporating broader, structured neural priors enhances the NBM's

predictive power for generalization behaviors, and that the NBM's predictions extend beyond purely exemplar-based similarity.

We found that after controlling for the BBM's predicted values using partial correlation, the NBM still exhibited unique predictive power (Fig. 4c; Pearson $R = 0.192$, 95% confidence interval = [0.024, 1], $BF_{10} = 2.372$, $T_{94} = 1.894$, $P_{\text{one-tailed}} = 0.031$), indicating the specific contributions of neural priors.

We found that the NBM significantly predicted generalization behavior for learning with familiar objects (reported above) but not for novel shapes (Fig. 4d; Pearson $R = 0.277$, 95% confidence interval = [-0.075, 1], $BF_{10} = 1.557$, $T_{22} = 1.350$, $P_{\text{one-tailed}} = 0.095$), although the difference between these two learning conditions was not significant ($z = 0.049$).

We further conducted the same analyses using linear mixed-effects models to control for the random effects of word concepts. Again, the results were consistent with those reported above (for all model comparisons, standardized coefficients > 0.110 , $Ps < 0.017$; see Section B4 in the Supplementary Information for details).

To summarize, in the VOTC, a region known for storing rich object knowledge, the NBM significantly predicted both the neural representations of newly learned word concepts and participants' generalization behavior in word concept learning with familiar objects (with rich priors). Such a Bayesian model with neural priors (1) outperforms neural mean models (i.e., neural models that do not incorporate broader structured priors) and prior-permuted control models, (2) provides predictive power beyond that of a Bayesian model using behavioral priors, and (3) has more advantages in explaining word concept learning with familiar objects than with novel shapes (i.e., weaker priors), especially in neural representation prediction. These results demonstrate the adequacy and necessity of involving brain regions that represent prior knowledge in word-concept learning, in a manner consistent with Bayesian inference mechanisms.

Word concept learning in the hippocampus and VMPFC/DMPFC

Surprisingly, the same NBM, when constructed using dendrograms derived from neural activity in the hippocampus, VMPFC, and DMPFC, failed to predict the neural representations of newly learned word concepts in these respective regions (Fig. 5a; $Ps_{\text{one-tailed}} > 0.199$) or participants' generalization behavior for familiar objects ($Ps_{\text{one-tailed}} > 0.300$).

We further considered the possibilities that these regions play a role in Bayesian word concept learning, but in ways different from those observed in the VOTC: (1) They may encode individual experiences, leading to sparse and distinct representations for each participant; (2) These regions may not encode structured priors themselves but instead rely on prior knowledge represented in cortical regions such as the VOTC; (3) They may be involved in tracking the dynamic updating of conceptual representations over the course of learning (see Theves et al.¹⁰).

First, the use of group-level priors assumes that the same object stimulus leads to similar neural representations across all participants. This assumption is often granted in the VOTC³⁹. However, for the other three ROIs, especially the hippocampus and VMPFC, it is possible that they encode individual experiences^{9,11–13,40,41}, potentially impairing the performance of the NBM when using group-level priors. Compared with the VOTC, the hippocampus did not exhibit lower cross-subject alignment, while the other two ROIs indeed showed lower alignment (Fig. S6b). Nonetheless, using individual priors did not produce robust NBM effects in the DMPFC or VMPFC. In the hippocampus, while the NBM had significant predictive efficacy, it was not significantly different from that of the NMM (Fig. S7 and Table S4). This indicated that the null findings for the hippocampus, VMPFC, and DMPFC cannot be attributed solely to the use of group-level pattern construction.

Second, is it possible that these regions participate in forming the neural representations of word concepts, but rely on prior knowledge

stored in other brain areas, such as the VOTC? To test this hypothesis, we performed RSA to compare the RDM of the posterior probability patterns from the VOTC-based NBM with the RDMs of the observed neural patterns of word concepts in the hippocampus, VMPFC, and DMPFC (Fig. 5b). A significant correlation between the two RDMs would suggest that the latter brain region contributes to new word representation formation by utilizing structured priors stored in the VOTC. The RSA results revealed that the VOTC-based NBM-model RDM did not significantly correlate with neural RDMs in the hippocampus, VMPFC, or DMPFC ($P_{\text{one-tailed}} > 0.201$; Fig. 5b and Table S5). The VOTC-based NBM-model RDM did significantly correlate with neural RDMs in the VOTC ($R_{\text{Fisher-Z}} = 0.156$, lower boundary of 95% confidence interval = 0.088, Cohen's $d = 0.894$, $BF_{10} = 92.148$, $T_{19} = 4.000$, $P_{\text{one-tailed}} = 3.83 \times 10^{-4}$), replicating the results in the above sections. That is, even considering prior building from the VOTC, neural activity patterns in the hippocampus, VMPFC, and DMPFC during word concept learning were not effectively predicted by the Bayesian model.

Third, is it possible that, while these brain regions do not directly predict learned concepts or generalization, they may be selectively involved in other roles of word concept learning, such as tracking and updating internal conceptual representations as learning unfolds¹⁰? Specifically, as each new exemplar is presented, the neural representation updates, reflected in changes in the posterior probability distribution across hypotheses in the NBM. These regions, being sensitive to learning, might track the magnitude of such updates. We conducted a parametric modulation analysis to test this possibility (Fig. 5c), examining whether the activation strength for the first word concept learning event (exemplar 1), the second word concept learning event (exemplar 2), and the third word concept learning event (exemplar 3) was a function of the degree of change in the posterior distribution across hypotheses. We constructed the VOTC-based NBM to compute the updating for each learning stage (Fig. 5c): For exemplars 2 and 3, the representation updating strength was operationally defined as dissimilarity (1-R) between the posterior probability pattern of hypotheses for the current exemplar and that of the previous exemplar; for exemplar 1, it was the dissimilarity (1-R) between the prior probability pattern and its posterior distribution. The parametric modulation results revealed that the VOTC-based NBM did not predict the activation strength in the hippocampus, VMPFC, or DMPFC ($P_{\text{one-tailed}} > 0.400$), but did so in the VOTC (Beta = 1.794, lower boundary of 95% confidence interval = 0.233, Cohen's $d = 0.444$, $BF_{10} = 2.275$, $T_{19} = 1.987$, $P_{\text{one-tailed}} = .031$; Fig. 5c and Table S6). Thus, there was no evidence that the hippocampus, VMPFC, or DMPFC tracked the neural pattern changes predicted by the Bayesian model.

Interestingly, while the above results showed that the NBM in the hippocampus, VMPFC, or DMPFC did not significantly predict the neural representations of words learned with familiar objects, when learning with novel shapes, the NBM incorporating priors derived from the hippocampus significantly predicted the new word neural representation in the hippocampus (Pearson $R_{\text{Fisher-Z}} = 0.020$, lower boundary of 95% confidence interval = 0.006, Cohen's $d = 0.534$, $BF_{10} = 4.399$, $T_{19} = 2.389$, $P_{\text{one-tailed}} = 0.014$; Fig. 5d and Table S7). This was not observed in the VMPFC or DMPFC ($P_{\text{one-tailed}} > 0.700$). Notably, the NMM also yielded significant predictions in the hippocampus and not in the VMPFC or DMPFC (hippocampus: Pearson $R_{\text{Fisher-Z}} = 0.021$, lower boundary of 95% confidence interval = 0.007, Cohen's $d = 0.593$, $BF_{10} = 6.970$, $T_{19} = 2.652$, $P_{\text{one-tailed}} = .008$; VMPFC/DMPFC: $P_s > 0.460$). There was no significant difference in predictive power between the NMM and the NBM (difference = 0.001, 95% confidence interval = [-0.007, 0.006], Cohen's $d = 0.053$, $BF_{10} = 0.238$, $T_{19} = -0.236$, $P_{\text{two-tailed}} = 0.816$). Partial correlation analysis revealed no unique predictive power for the NBM beyond the NMM (Pearson $R_{\text{Fisher-Z}} = 0.009$, lower boundary of 95% confidence interval = -0.001, Cohen's $d = 0.354$, $BF_{10} = 1.247$, $T_{19} = 1.581$, $P_{\text{one-tailed}} = 0.065$) or for the

NMM beyond the NBM (Pearson $R_{\text{Fisher-Z}} = 0.002$, lower boundary of 95% confidence interval = -0.008, Cohen's $d = 0.066$, $BF_{10} = 0.294$, $T_{19} = 0.293$, $P_{\text{one-tailed}} = 0.386$) in the hippocampus.

To summarize, for the hippocampus, VMPFC, or DMPFC, we did not observe evidence for their involvement in learning with rich priors using the NBM: the NBM failed to predict their neural representations of new words; using individual priors from the corresponding ROIs failed to yield robust predictions in the VMPFC/DMPFC and had no advantage over the NMM in the hippocampus; using neural priors from the VOTC failed to yield significant results; and there was no evidence that they were sensitive to the updating of new word concepts during learning. By contrast, for word concept learning with weak priors, the NBM did predict the new word representations in the hippocampus, yet with comparable predictive power to that of the NMM; both models showed stronger effects for predicting neural representations of word concepts based on novel shapes than for those based on familiar objects (Table S7). That is, hippocampal neural representations of new words are a function of the associated novel exemplars, without stable evidence of prior contributions.

Different patterns of the VOTC and hippocampus in learning with rich versus weak priors

In the above ROI analyses, we observed different patterns regarding the effects of the NBM in learning with rich or weak priors across ROIs: in the VOTC, which has been implicated in object representation, the NBM had greater predictive power in learning with familiar objects than in learning with novel shapes; in the hippocampus, which has been previously implicated in learning concepts from novel and/or artificially designed stimuli, the NBM stably predicted neural representations of word concepts learned with novel shapes, but not with familiar objects.

Direct statistical analyses confirmed these visual inspections. A significant two-way interaction effect was observed (VOTC/hippocampus $\times$ familiar objects/novel shapes): $F_{(1,19)} = 10.304$, $\eta^2 = 0.352$ (95% confidence interval = [0.137, 0.747]), $P_{\text{two-tailed}} = 0.005$ (Fig. 5e). Further simple effect analyses revealed that the NBM had better predictive power for learning with familiar objects than with novel shapes in the VOTC (Fig. 3a, top), and better predictive power for learning with novel shapes than with familiar objects in the hippocampus (difference = 0.015, 95% confidence interval = [0.001, 0.028], Cohen's $d = 0.504$, $BF_{10} = 1.794$, $T_{19} = 2.254$, $P_{\text{two-tailed}} = 0.036$).

Whole-brain searchlight results in predicting new word neural representations (with rich priors)

The whole-brain searchlight analysis employed the same methods as the ROI analysis and revealed similar results to those in the VOTC (Fig. S8 and Tables S8 and S9). The NBM predicted the new word neural representations in widely distributed regions encompassing the bilateral fusiform gyrus, lateral occipitotemporal cortex, inferior parietal sulcus, supplementary motor area, precentral gyrus to middle frontal gyrus, and inferior frontal gyrus (voxel-wise $P_{\text{one-tailed}} < 0.001$ using cluster-level family-wise error (FWE)-corrected $P < 0.05$), which were consistently found to represent object knowledge^{35,37,38,42,43}. The most robust effects were obtained in the bilateral occipitotemporal fusiform gyrus (within the VOTC), where the NBM showed greater predictive power than both the NMM and prior-permuted control models, predictive power beyond that of the BBM, and greater predictive power in learning with familiar objects than with novel shapes. No significant result was found in the hippocampus, VMPFC, or DMPFC. These results, together with the ROI results, converge to demonstrate that the brain regions representing object knowledge support word concept learning via Bayesian inference mechanisms. For more detailed information, see the Section B5 in the Supplementary Information.

Table 2 | Results of predicting human generalization behavior using LLMs and the Bayesian learning model

Learning condition	Models	<i>R</i>	95CI (lower boundary)	<i>BF</i> ₁₀	<i>T</i>	<i>P</i>	Compared with the Bayesian learning model	
							<i>z</i>	<i>P</i>
Overall	GPT-4o	0.677	0.586	3.863E + 14	9.998	1.02E-17	4.990	6.04E-7
	Qwen2.5-VL	0.314	0.171	1.414E + 2	3.588	2.43E-4	8.326	8.36E-17
	Bayesian learning model*	0.897	0.863	2.989E + 39	22.083	5.01E-44	–	–
Familiar objects	GPT-4o	0.649	0.539	1.452E + 10	8.269	4.34E-13	4.889	1.02E-6
	Qwen2.5-VL	0.261	0.096	10.542	2.621	0.005	6.941	3.90E-12
	Bayesian learning model	0.905	0.869	3.212E + 32	20.678	4.76E-37	–	–
Novel shapes	GPT-4o	0.795	0.620	4.762E + 3	6.140	1.76E-6	1.349	0.177
	Qwen2.5-VL	0.468	0.148	7.531	2.485	0.011	2.798	0.005
	Bayesian learning model	0.892	0.791	6.772E + 5	9.279	2.31E-9	–	–

95CI: 95% confidence interval.

Note. The degrees of freedom were 118, 94, and 22 for overall learning, familiar-object learning, and novel-shape learning, respectively. The *P* value for “Compared with Bayesian learning model” contrast is obtained via Pearson and Filon’s *Z*-test (two-tailed). All other tests were one-tailed. No adjustment was made for multiple comparisons. *The Bayesian learning model’s predictions were derived by regressing observed generalization behavior onto the predictions of both the VOTC-based NBM and the BBM using a general linear model, via a leave-one-trial-out approach. Source data are provided as a Source Data file.

Bayesian model predicts human concept learning behavior better than LLM

The recent generation of LLMs, while not explicitly implementing the Bayesian model in training, exhibit human-like language processing skills^{44–46}. This leaves open the question of whether LLMs, endowed with extremely rich, large-scale semantic representations, would, by virtue of that richness alone, exhibit human-like word concept learning behavior within the same learning scenario. To test this, we asked LLMs to perform the same learning task as humans and compared the LLMs’ predictions with those of the Bayesian learning model in accounting for human word concept learning behavior. As human word concept learning scenarios require the simultaneous input of visual images and language instructions, we tested two state-of-the-art, multimodal LLMs to align as closely as possible with the human paradigm: GPT-4o-2024-11-20⁴⁷, Qwen2.5-VL⁴⁸. We prompted these models to perform the same task as in Experiment 2 (learning experiment), conducted in independent chat sessions with 20 repetitions per word-concept (following a procedure similar to Strachan et al.⁴⁹), with temperature = 1.0 and top-*P* = 1.0 to sample from the models’ original output distributions (Fig. 6a). Each model’s predicted probability of a probe belonging to a given word-concept was calculated as the proportion of positive responses over 20 iterations. The results of GPT-4o and Qwen2.5-VL significantly correlated with human word concept learning generalization behavior (combining both familiar and novel objects; for GPT-4o, *R* = 0.677, 95% confidence interval = [0.586, 1], *BF*₁₀ = 3.863 × 10¹⁴, *T*₁₁₈ = 9.998, *P*_{one-tailed} = 1.02 × 10^{−17}; for Qwen2.5-VL, *R* = 0.314, 95% confidence interval = [0.171, 1], *BF*₁₀ = 1.414 × 10², *T*₁₁₈ = 3.588, *P*_{one-tailed} = 2.43 × 10^{−4}; Fig. 6b).

Since the VOTC-based NBM provided additional predictive power beyond the BBM in predicting generalization behavior, we calculated the best predictions of the Bayesian learning model by combining the NBM and the BBM, and compared this combined model to the LLMs. We fitted a general linear model to predict the observed human generalization behavior from the predictions of both the NBM and the BBM via a leave-one-trial-out approach. The Bayesian learning model is highly correlated with human behavior (*R* = 0.897, 95% confidence interval = [0.863, 1], *BF*₁₀ = 2.989 × 10³⁹, *T*₁₁₈ = 22.083, *P* = 5.01 × 10^{−44}) and significantly more strongly than the LLMs (*z*s > 4.990, *P*s < 0.001; Table 2). These advantages persisted when analyzing learning with familiar objects only (*z*s > 4.889, *P*s < 0.001). In learning with novel shapes, this advantage was absent between the Bayesian learning model (*R* = 0.892, 95% confidence

interval = [0.791, 1], *BF*₁₀ = 6.772 × 10⁵, *T*₂₂ = 9.279, *P* = 2.31 × 10^{−9}) and GPT-4o (*R* = 0.795, 95% confidence interval = [0.620, 1], *BF*₁₀ = 4.762 × 10³, *T*₂₂ = 6.140, *P* = 1.76 × 10^{−6}; difference, *z* = 1.349, *P* = 0.177), yielding a significant two-way interaction (Bayesian learning model/GPT × familiar objects/novel shapes; *z* = 19.878, *P* = 6.31 × 10^{−88}). That is, the tested LLMs do not simulate human word concept learning behavior as well as the Bayesian learning model, especially when learning with rich priors.

Discussion

To understand how the human brain learns word concepts based on prior knowledge, we developed and tested the NBM. This model is derived from the hypothesis that learning word concepts involves Bayesian inference that incorporates not only the representations of directly associated exemplars, but also the broader, structured neural priors. The results showed the advantages of the NBM over control models that lack such structured neural priors and over learning conditions that lack rich priors, in predicting both neural representations in the VOTC and generalization behavior for learning with familiar objects. Meanwhile, hippocampal activity was not associated with learning with familiar objects, but was associated with learning with novel shapes (i.e., weak priors). We discuss these key findings in turn below.

In learning with familiar objects, the NBM successfully predicted neural representations of word concepts in the VOTC, along with other regions commonly observed to represent object knowledge. The NBM considers the prior representational structure of not only the associated exemplars when forming the new word representation. It not only better predicted the neural representations of the learned concepts, but also significantly predicted the subsequent generalization behavior—the probability that a probe object belongs to the learned word. That is, the Bayesian computational model developed to account for the behavioral patterns of concept learning, previously criticized for not being neurally grounded^{50–52}, indeed offers a computational framework for neural mechanisms that support learning with existing representations in semantic memory. This is in line with the hypothesis that word concept learning may occur via a “fast mapping” mechanism, integrating new information into neocortical memory networks⁵³, especially when the new information is consistent with prior information^{40,41}. New word concepts may directly update the existing neural representations in the long-term (semantic) memory store^{40,41,53}.

In this paradigm, “word concept learning” refers to learning an extension for a novel label from exemplars and generalizing that label to new items, rather than merely memorizing word–image pairs. This process instantiates conceptual meaning for the label. The finding that priors derived from the VOTC successfully predict such learned conceptual representations indicates that the neural representational geometry in the ventral visual cortex provides effective priors for generalization in this task. However, the fact that it occurs in the VOTC, a canonical high-level visual area, does not imply that learning is “perceptual” and not semantic. Indeed, the nature of representation in the VOTC is extremely rich and high-dimensional (e.g., Bi et al.³⁵; Du et al.⁴⁴; Peelen and Downing⁵⁴). We simply interpret the VOTC result as evidence that word concept learning entails a Bayesian learning mechanism over the structured object neural representations here, while remaining agnostic about the precise feature content driving this learned word conceptual structure.

Note that for novel-shape concepts, the NBM did not significantly predict generalization behavior. This should not be overinterpreted as evidence that weak priors do not support Bayesian concept learning: the novel-shape condition comprised fewer trials (3 words; 24 trials), limiting the statistical power of the analysis; the familiar-object and novel-shape generalization effects did not show a significant difference. We therefore treat this pattern as suggestive and as motivation for future designs with increased sampling of weak-prior concepts.

When learning the association between a word token and visual stimuli without rich priors (novel shapes), the learned neural representation in the hippocampus is significantly explained by the composite of the neural representations of only the associated exemplars themselves (i.e., simple average), with additional consideration of prior structure adding no explanatory power (i.e., the NBM was not superior to the NMM). This result is in line with the literature on associative learning, showing that hippocampal responses (fMRI responses or new neuronal response modulations) are sensitive to learning specific associations of objects^{13,14,17,20,55}, or specific combinations of attributes^{9,11,12}, which are inherently more episodic in nature. Even using neural priors from the VOTC failed to make predictions in the hippocampus. These findings corroborate the hypothesis that learning word concepts with rich priors may bypass the hippocampus, or at least involve the hippocampus in a way that differs from classical associative learning, again consistent with the postulation of the “fast mapping” mechanism^{40,41}. How can these findings be reconciled with the classical evidence that hippocampal lesions lead to semantic learning deficits, as seen in cases such as patient H.M.^{56–58}? We speculate that the deficits stem from the episodic aspects of word concept learning, such as the need to remember specific word forms, which were not assessed in the present task.

Relatedly, when testing the different patterns regarding the effects of the NBM in learning with rich or weak priors, we observed stronger effects for familiar objects than for novel shapes in the VOTC, and vice versa in the hippocampus, which was corroborated by a significant interaction effect (VOTC/hippocampus $\times$ familiar objects/novel shapes). These results provide *prima facie* evidence for a double dissociation. However, a definitive conclusion about the double dissociation for learning with rich and weak priors requires caution, given that the test of interaction was conducted after observing the two simple effects, which carry certain inferential limitations. Additionally, the effects of the NMM in the hippocampus had predictive power comparable to that of the NBM. Thus, the specificity of the hippocampal effects regarding the role of priors remains unclear in our data. What we can conclude, nonetheless, is that the hippocampus did not show strong evidence for the NBM in learning with rich priors, unlike the VOTC.

We did not observe significant effects of either NBM or NMM on learning word concepts with familiar objects or novel shapes in the medial prefrontal cortex, in predicting either learned neural

representations or concept representation updating as a function of changes in the VOTC representations during learning. These null results contrast with positive findings regarding the involvement of the medial frontal cortex in learning categories or associations^{10–12,19,20,55}. In addition, other medial temporal regions, such as the perirhinal cortex, have also been shown to encode specific learned associations or concepts^{13,15,16,59–64}, but did not exhibit significant effects in our whole-brain analyses. The negative findings are difficult to interpret. One possibility is that these regions still engage in word concept learning and establish relationships with associated stimulus representations, but not in the ways we assessed here, or that they are sensitive to associative representations within a broader context than the limited object space sampled in our experiment. The exact computational role of these regions in word concept learning remains to be understood.

Our findings contribute to ongoing efforts to bridge computational models of learning and reasoning with their underlying neural mechanisms. While Bayesian models have successfully captured behavioral patterns in word concept learning, they are often critiqued for lacking mechanistic grounding—defining optimal solutions without specifying how such computations are implemented in the brain⁶⁵. Here, our finding that neural representations in the VOTC, but not in the hippocampus, predicted both the concept representations and behavioral generalization suggests that probabilistic inference may be implemented in domain-specific, long-term memory storage cortical circuits. This suggests that priors in Bayesian word concept learning may be instantiated in the representational topology of cortical systems, and that word concept learning may proceed through reshaping this topology. More broadly, rational models such as Bayesian inference are sometimes criticized as overidealizing human cognition, given evidence that human behavior is frequently biased and variable. By grounding the priors and inferences underlying concept learning in distributed and graded neural codes, our approach opens a path toward understanding how variability in the neural priors underpins the individual differences in word concept learning and generalization.

This framework also has implications for the intricate relationship between episodic and semantic memory, which has long been recognized^{66–70}. Recent evidences show that these two types of memory share large overlapping neural correlates (as reviewed by Renoult et al.⁶⁸), that the relationships between stimuli in semantic memory can influence performance on episodic memory tasks^{21–23}, and that impairments in one can affect the function of the other^{71–74}. Our findings contribute to this understanding of the dynamic interplay between episodic and semantic memory. Specifically, learning in the context of (1) associating a word token with familiar objects and (2) instructions that encourage categorization and generalization, is supported by Bayesian inference operating on prior (semantic) long-term neural representations in the cortex. This contrasts with reports that learning arbitrary associations of specific stimuli primarily involves the hippocampus. It remains unclear whether these two factors (word-like token; task context) are necessary for the observed effects. Developmental research has highlighted the specific roles of verbal cues, reporting that children older than 6 months use new word cues with phonological patterns of their native language to learn the category based on associated objects (and to generalize accordingly), but not other cues, such as sounds of different linguistic groups, scrambled speech, or animal sounds^{75–77}. In addition, developmental studies have shown that word concept learning can occur in a one-shot manner, provided that a specific context is given when other prior cues are available, even without explicit categorization instruction or task (i.e., the fast mapping scenario^{78,79}). Whether neural mechanisms implementing Bayesian inference can explain these developmental behavioral observations is an important question for future research.

The target neural model was inspired by and developed from a Bayesian inference model for human word concept learning in the real world—linking words with external referents and nonverbal experiences. The recent generation of LLMs, despite not implementing such Bayesian mechanisms during training, has nonetheless achieved unprecedented success in human-like language processing, capturing internal representations with structures similar to those of human concepts, as measured both behaviorally and neurally^{44,80}. However, despite the billions of parameters ($\geq 7\text{B}$) and the super-rich prior knowledge in these LLMs obtained during pre-training^{47,48}, these models do not simulate human word concept learning behavior as well as the Bayesian learning model, especially when learning with rich priors. This indicates that semantic richness alone is insufficient to produce human-like word-concept learning in the absence of an appropriate inference mechanism that supports structured generalization from limited data, highlighting a distinction between possessing structured knowledge and deploying that knowledge to guide learning. Addressing this gap could inform the development of language models that better link prior experience to word concept learning.

To conclude, we investigated how the brain supports learning word concepts, which entails establishing (associative) mappings between words and exemplars, guided by long-term prior knowledge, and forming a conceptual space that supports generalization. The results revealed a Bayesian computational mechanism underlying neural representations in classical object representation regions such as the VOTC. This contrasts with learning with novel stimuli, for which neural representations in the hippocampus can be predicted by specific associated stimuli. These findings open further avenues for understanding concept learning in the broader context of episodic and semantic learning/representations.

Methods

Ethics approval

All protocols and procedures of the current study were approved by the local research ethics committee at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, China (Protocol ICBIR_A_0115_001). Each participant read and signed an informed consent form before participating in the experiments. All experiments were conducted in accordance with the Declaration of Helsinki and all relevant ethical regulations.

Participants

The participants were all right-handed native Mandarin Chinese speakers. None of them had a history of psychiatric or neurological disorders, nor had they sustained a head injury. The sample size was 10 (8 women; mean age, 21.20 years; SD, 1.14 years) for the pilot experiment, and 20 (13 women; mean age, 21.45 years; SD, 2.06 years) for Experiment 1, Experiment 2, and the semantic distance-judgment task. Each participant in the pilot experiment received 30 RMB. Each participant in Experiments 1 and 2 received 230 RMB. Another 20 participants (12 women; mean age, 21.80 years; SD, 1.99 years) participated in the visual-shape distance-judgment task for familiar objects, with each participant receiving 30 RMB.

Sample sizes were chosen in reference to prior fMRI work using comparable representational analyses and stimulus domains, as well as behavioral studies of Bayesian word learning. For instance, for multivoxel pattern analysis (MVPA) of object neural representations, data from 12 participants were sufficient to distinguish the object shape knowledge of 33 different artifacts ($T_{11}=2.50$, Cohen's $d=0.722$; see Fig. 3b in Peelen et al.⁸¹). In the behavioral study¹, data from 22 participants were sufficient to predict the generalization behavior of 12 words (96 trials) using the BBM ($R=0.89$, Cohen's $d=3.903$). We also reported the effect sizes in our analyses (Tables S2–S7).

Designs and procedures

The experimental materials consisted of 58 grayscale object pictures (300×300 pixels), including 13 animals, 13 famous human faces, 19 human-made artifacts, and 13 novel shapes. Within each domain, 8 objects were used as probe objects, and the remaining objects served as exemplars during learning. All stimuli are shown in Fig. S1. The pilot experiment followed the same procedure as Experiment 2 but was conducted outside the scanner to confirm the feasibility of the fMRI session.

In Experiment 1, participants were asked to view each object picture and to think about its meaning while undergoing fMRI scanning. Each picture was presented for 1 s, followed by a fixation interval. Each fixation lasted 1–3 s (mean = 2 s). When the fixation point turned red, participants needed to judge whether the upcoming object was similar to the previous one and to press a key to indicate their judgment. This task included ten runs, each lasting 4 min 28 s, and employed an event-related design. Each run included 58 object pictures and 11 object pictures as catch trials. Fifteen new words were also included as filler trials. The number and order of trials for different stimuli were counterbalanced across runs and participants.

In Experiment 2, participants performed a word concept learning task in which they were asked to help a friend who spoke another language to select desired objects. In each trial, this friend would show either one or three exemplars in turn, each labeled with a word token lasting 2 s. Subsequently, the word token appeared alone again for 3 s, followed by a probe object lasting 2 s. The exemplars, the word token, and the new probe object were separated by fixation intervals. Each fixation lasted 0.5–1.5 s (mean = 1 s). Participants needed to learn the new word's concept based on the exemplars and then press a key to indicate whether the new probe object was the one their friend wanted. We manipulated the number and similarity of exemplars, resulting in three different new words in each domain: the first word token with a single exemplar, the second word token with three exemplars of high similarity, and the third word token with three exemplars of low similarity (Fig. S1). Each domain included eight probe objects that were not used as exemplars (Fig. S1). There were also three other artifact words that differentiated between visual and functional similarities (exemplars that were highly similar in only visual aspects, only functional aspects, or both aspects), resulting in 15 different word concepts in total (12 with familiar-object exemplars and 3 with novel-shape exemplars). This task included 4 runs of 8 min 28 s each. Each word token had two trials in each run, with the number and order of trials for different words counterbalanced across runs and participants. Each run in each experiment started with a fixation lasting 10 s and ended with a fixation lasting 6 s.

After fMRI scanning, participants were asked to perform an online semantic distance-judgment task using a multi-arrangement paradigm⁸² via NAODAO (<https://www.naodao.com>). In this paradigm, participants dragged and dropped the 58 objects into a circular array on a computer screen, arranging them spatially close together or far apart according to the semantic distances between the objects. This task lasted 30 minutes. The visual-shape distance-judgment task followed the same procedure, except that participants were asked to arrange the 45 familiar objects according to their visual-shape distance.

Finally, we prompted two LLMs—GPT-4o-2024-11-20 and Qwen2.5-VL-7B-Instruct—to perform the same word concept learning task as in Experiment 2. For a given trial (e.g., deciding whether an axe belongs to “Leca” given the exemplars spoon, hammer, and eyeglass), the exemplars were presented in three separate messages, mirroring the human experiment (see Fig. 6a for prompts). Each trial was repeated 20 times in an independent chat session, with temperature = 1.0 and top- $P=1.0$ to sample from the models' original output distributions. The generalization probability for a trial was averaged across the 20 iterations.

Image acquisition and preprocessing

All functional and structural MRI data were collected on a Siemens 3 T Prisma scanner with a 64-channel head-neck coil at the Imaging Center for Brain Research, Beijing Normal University. Functional data were acquired with a simultaneous multi-slice echoplanar imaging sequence supplied by Siemens (64 axial slices, repetition time [TR] = 2000 ms, echo time [TE] = 30 ms, multi-band factor = 2, flip angle [FA] = 90°, field of view [FOV] = 208 mm × 208 mm, matrix size = 104 × 104, slice thickness = 2 mm, gap = 0.2 mm, and voxel size = 2 mm × 2 mm × 2 mm). A high-resolution 3D T1-weighted anatomical scan was acquired using the magnetization-prepared rapid acquisition gradient echo sequence (192 sagittal slices, TR = 2530 ms, TE = 2.98 ms, inversion time = 1100 ms, FA = 7°, FOV = 224 mm × 256 mm, matrix size = 224 × 256, interpolated to 448 × 512, slice thickness = 1 mm, and voxel size = 0.5 mm × 0.5 mm × 1 mm).

The fMRI data were preprocessed using the Statistical Parametric Mapping software (SPM12; <http://www.fil.ion.ucl.ac.uk/spm/>) and the advanced edition of DPARSF V4.3⁸³ implemented in DPABI V3.0⁸⁴. For the preprocessing of the task fMRI data, the first five volumes of each functional run were discarded to reach signal equilibrium. Slice timing and 3-D head motion correction were performed. Subsequently, a mean functional image was obtained for each participant, and the structural image of each participant was coregistered to the mean functional image. Thereafter, the structural image was segmented using a unified segmentation module⁸⁵. Next, a custom, study-specific template was generated by applying diffeomorphic anatomical registration through exponentiated Lie algebra (DARTEL⁸⁶). The parameters obtained during segmentation were used to normalize the functional images of each participant into the Montreal Neurological Institute space by applying the deformation field estimated by segmentation. The functional images were subsequently spatially smoothed using a 6-mm full-width-half-maximum Gaussian kernel for univariate parametric analysis but not for MVPA.

Data analysis

For the behavioral analyses and results, please see the Supplementary Information (Section A). The fMRI data analyses were all conducted using SPM12 and custom R scripts⁸⁷ unless specifically stated.

First-level analysis. As a sanity check, a general linear model analysis was performed to test the domain-specific activations for each participant in Experiment 1. Specifically, four regressors of interest were modeled, corresponding to pictures of animals, human faces, artifacts, and novel shapes, respectively. The filler trials and the catch trials were each modeled as a separate nuisance regressor. For the three familiar-object domains (animals, faces, artifacts), domain-specific activations were calculated as the contrast of one domain against the other two. For the novel-shape domain, its domain-specific activation was calculated as the contrast of novel shapes against all three familiar-object domains.

For model construction and prediction, a GLM analysis was performed to explore the fixed effect of each object for each participant in Experiments 1 and 2, respectively. In Experiment 1, each of the 58 objects was modeled as a regressor of interest. Each of the filler trials was modeled as a nuisance regressor. All catch trials were modeled as one nuisance regressor. In Experiment 2, for each new word, two regressors of interest were modeled, one for the exemplars and the other for the single presentation of the new word. All probe objects were modeled as a single nuisance regressor. In both experiments, six head motion parameters obtained from head motion correction were also included as nuisance regressors, and a high-pass filter (128 s) was used to remove low-frequency signal drift for each run. A group-level averaged *t* value map across the participants for each of the 58 objects in Experiment 1 was estimated after GLM analyses; these maps served as the neural priors. In Experiment 2, a *t* value map combining the

exemplars and the single new word was estimated for each word and participant via the GLM; this map was taken as the neural representation of the word concept.

The group-level *t*-value maps for each of the 45 familiar objects were first utilized to validate the advantage of neural priors, i.e., capturing the diverse types of object representations distributed across the brain. For each region, the activation pattern of each familiar object was extracted; from these patterns, the dissimilarity (1-R) between each object pair was calculated, yielding a neural RDM. Meanwhile, behaviorally rated semantic and shape-based RDMs for the familiar objects were derived from the distance-judgment tasks, providing two targeted model RDMs. Another pixel-based RDM was also considered as a low-level visual control variable; it was generated by calculating the dissimilarity (1-R) of grayscale pixel values between each pair of pictures. For each region, partial correlation analyses were conducted to calculate the unique correlation (Spearman's rho) between the neural RDM and each target model RDM, while controlling for the other target model RDM and the pixel-based RDM.

Because rapid event-related designs can be sensitive to the choice of first-level estimation procedure for item-specific patterns⁸⁸, we compared the Least Squares-All (LS-All) approach we used (see above), which estimates all trials in a single regression, and the Least Squares-Separate (LS-S) approach, which estimates each trial in a separate regression. We assessed which approach could (1) better capture the unique semantic and visual-shape information (RSA-based validation) within the ventral visual cortex and (2) achieve higher ISC in representational similarity patterns, given that the NBM depends on priors built from reliable item-wise representational geometry. In our data, the LS-All approach overall yielded statistically comparable or even stronger representational sensitivity and ISC than the LS-S approach (Fig. S6). We therefore used the estimates from the LS-All approach for the main analyses and have also uploaded the whole-brain data obtained using both the LS-All and LS-S approaches to facilitate further exploration (<https://doi.org/10.17605/OSF.IO/WRT9S>).

Model construction, prediction, and comparison analyses were conducted using both whole-brain searchlight MVPA (with a 5 × 5 × 5 cube size centered on each voxel) and ROI-based MVPA. The whole-brain searchlight analysis was conducted within the group-based gray matter mask. To obtain the mask, each participant's normalized structural image was segmented into different tissue classes. The resulting gray-matter probabilistic images were resliced to the functional image resolution, averaged across participants, and thresholded at 0.25 to produce a binary mask for searchlight mapping. All whole-brain and ROI-level prediction and comparison procedures were implemented with custom scripts in R.

The Harvard-Oxford Atlas was used to define the VOTC, hippocampus, and VMPFC (frontal medial cortex). Specifically, the VOTC was defined as the union of the following bilateral regions: inferior lateral occipital cortex, occipital fusiform, occipitotemporal fusiform, inferior occipitotemporal cortex, and middle occipitotemporal cortex. The DMPFC was defined as the sum of 5 × 5 × 5 voxel cubes centered on peak coordinates identified in previous studies related to learning associations based on prior knowledge. These coordinates included [2, 46, 0] in van Kesteren et al.²⁰, [-2, 40, 2] in van Kesteren et al.⁵⁵, and [-6, 34, 12] in van Kesteren et al.¹⁹.

NBM. For a given region, the NBM was constructed from the group-averaged *t*-value patterns for each of the 58 objects in Experiment 1. A hierarchical clustering analysis was performed on the dissimilarities among the *t*-value patterns of the 58 objects to construct the prior knowledge dendrogram using *complete linkage* (default in R), which merges clusters based on the maximum distance between their members. Each node in the dendrogram represented a possible hypothesis. The prior, likelihood, and posterior for each node, and the

generalization behavior for each probe object, were calculated as follows:

Prior probability computation. Following earlier work, the prior probability of a node was defined as the height difference between the node and its parent:

$$p(h) \propto \text{Height}[\text{parent}(h)] - \text{Height}(h) \quad (1)$$

Because of the inherent noise in fMRI data, objects from different domains could be clustered together (i.e., a node might contain stimuli from multiple domains). Therefore, following Xu and Tenenbaum¹, we introduced two modifications to enable the use of all nodes in the dendrogram. First, node heights were scaled to the range 0–0.5, where 0 corresponds to the lowest leaf and 0.5 to the highest root node. Second, for the highest node, its prior probability was set to 0.5, assuming the existence of a virtual parent node with height 1, consistent with the largest height in Xu and Tenenbaum¹.

Likelihood computation. In the word concept learning task (Experiment 2), each new word was associated with one or three exemplar objects. Given the exemplars, the likelihood for a node was calculated as the reciprocal of the node height raised to the power of n (i.e., *size principle* in Xu and Tenenbaum¹), where n is the number of exemplars (X) of the current word contained in that node (Eq. (2)):

$$p(X|h) \propto \left[\frac{1}{\text{Height}(h) + \varepsilon} \right]^n \quad (2)$$

If a node contains none of the exemplars, its likelihood is zero. A small constant value (ε) was added to the height to avoid the likelihood probabilities being infinite at the lowest nodes.

Posterior probability computation. After the likelihoods were calculated, the posterior for each node was calculated via Bayes' theorem (Eq. (3)):

$$p(h|X) = \frac{p(h)p(X|h)}{\sum_{h' \in H} p(h')p(X|h')} \quad (3)$$

where H represents the set of all hypotheses.

Generalization probability computation. The probability that a probe object (y) belongs to the current new word (W) is calculated by summing the posterior probabilities of all nodes that simultaneously contain both the probe object and at least one exemplar (Eq. (4)):

$$p(y \in W|X) = \sum_{h \supset y, X} p(h|X) \quad (4)$$

Predicted neural representation computation. The predicted neural representation of a learned concept (e.g., “Leca”) was computed as a weighted sum of the neural activity patterns (from Experiment 1) of all objects in the same domain, with weights given by each object's generalization probability for the concept. Figure 2 shows the flowchart of predicting the neural representations of a new word (e.g., “Leca”).

NMM. The NMM serves as a control model for the NBM, which predicts the learned word concept representation by simply averaging the neural activation patterns of its associated exemplars (see below) and does not incorporate structured priors but still uses neural signals. This prediction is thus sensitive to perceptual-pictorial similarity between objects, and comparing its performance against that of the NBM allows us to rule out the possibility that the NBM's predictions are driven solely by such low-level pictorial information.

Predicted neural representation computation. Under the NMM, the neural representation of a new word is predicted as the average of the t value patterns of its associated exemplars. For example, for “Leca”, the NMM's predicted representation was the averaged t value pattern of the spoon, hammer, and eyeglass (observed in Experiment 1).

Generalization probability computation. For a probe object (e.g., the axe), its generalization probability of belonging to one word concept (e.g., “Leca” learnt by referring to a spoon, a hammer, and an eyeglass) was calculated as the Pearson correlation between the t -value pattern of the probe object (e.g., the axe) in Experiment 1 and the NMM's predicted pattern for the word (e.g., the average t value pattern of the spoon, hammer, and eyeglass, also from Experiment 1).

Prior-permuted NBM. The prior-permuted NBM was constructed and applied using the same steps as the NBM, except that object identities were randomly permuted before constructing the prior dendrogram.

BBM. The BBM was constructed and applied analogously, with the prior dendrogram derived from the behavioral semantic distance-judgment task.

Model comparison. For each model, predictive power for a given word's neural representation was quantified as the Fisher-transformed Pearson correlation between the predicted and observed neural patterns, separately for each participant. For each participant and model, Fisher- z values were averaged separately for words referencing familiar objects ($n = 12$) and for those referencing novel shapes ($n = 3$). We conducted paired-samples t tests to compare the predictive power of the NBM and NMM for familiar-object learning, and to compare the predictive power of the NBM between familiar-object and novel-shape learning. Partial correlation analyses were used to test whether the NBM provided unique predictive power beyond the NMM and BBM. For each participant and word, the Fisher-transformed Pearson correlation between the NBM-predicted and observed neural patterns was computed, with the predicted neural pattern of the NMM or BBM controlled for, respectively. The resulting Fisher- z values were averaged across familiar-object word concepts and submitted to a one-sample t test. A permutation test assessed whether the NBM outperformed the prior-permuted NBM, in which the object-to-pattern mapping was randomly permuted 100 times. A null hypothesis distribution was constructed via a bootstrap procedure ($n = 10,000$ iterations). In each iteration, for each participant, one of the 100 permuted versions of each word was randomly drawn (with replacement), and the correlation between the predicted and observed neural patterns was computed and Fisher- z -transformed. The group-level averaged Fisher- z values computed in each bootstrap iteration formed the null distribution. To provide a quantitative measure of the magnitude across regions, we calculated the SES as $\frac{x - \mu}{\sigma}$, where x is the observed mean, μ is the mean of the null distribution, and σ is its standard deviation⁸⁹. A one-tailed P value was estimated by fitting a standard normal distribution to the null distribution.

The predictive power of each model for generalization behavior was assessed as the Pearson correlation between the predicted and observed generalization probabilities across all trials (i.e., all word–probe combinations). Pearson and Filon's Z -test was used to compare the predictive power of the NBM and NMM. Partial correlation analysis tested whether the NBM provided unique predictive power beyond the NMM and BBM. A permutation test was performed (1000 iterations) to evaluate whether the NBM outperformed the prior-permuted NBM. The correlation coefficients from these permutations formed the null distribution, from which the SES for the observed correlation was computed. The one-tailed P value was estimated as described for the neural prediction analysis. A Fisher- z test

compared the NBM's predictive power between familiar-object and novel-shape learning.

Whole-brain-level analysis. For the NBM, NMM, and BBM, the searchlight analysis resulted in a Fisher-z map for each word concept and participant. Two Fisher-z maps of the NBM were also obtained for each word and participant in the partial correlation analysis, when the predicted patterns of the NMM and BBM were controlled for. All maps were then smoothed using a 6 mm FWHM Gaussian kernel, and the model comparisons described above were performed using SPM12. Multiple comparison corrections were conducted using cluster-level FWE correction ($P < 0.05$) as implemented in SPM12 (voxel-wise $P < 0.001$). For the whole-brain-level permutation test, we performed it in each of 96 cortical regions in the Harvard-Oxford Atlas instead of at the voxel level due to computational constraints, and multiple comparisons were corrected across the entire brain regions using the false discovery rate (FDR) correction algorithm ($q < 0.05$). For the generalization behavior analysis, for each trial in familiar-object learning, the NBM predicted a generalization probability map. These maps were smoothed using a 6 mm FWHM Gaussian kernel and then entered into a regression analysis implemented in SPM12. Multiple comparison corrections were conducted using cluster-level FWE correction ($P < 0.05$) as implemented in SPM12 (voxel-wise $P < 0.001$). The brain results were projected onto the MNI brain surface for visualization with BrainNet Viewer⁹⁰.

Reporting summary

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

Data availability

The data that support the findings of this study are available in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>) under the CC BY 4.0 License. Source data are provided with this paper (<https://doi.org/10.17605/OSF.IO/WRT9S>). The stimuli of novel shapes were originally created by authors themselves, with the raw image files uploaded in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

Code availability

Custom code that supports the findings of this study is available from Open Science Framework (<https://doi.org/10.17605/OSF.IO/WRT9S>).

References

- Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007).
- Waxman, S. R. & Markow, D. B. Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.* **29**, 257–302 (1995).
- Lupyan, G. Labels facilitate learning of novel categories. In Cangelosi A., Smith A. D. M., Smith K. R. (Eds.), *The Evolution of Language* 190–197 (World Scientific, 2006).
- Lupyan, G., Rakison, D. H. & McClelland, J. L. Language is not just for talking: redundant labels facilitate learning of novel categories. *Psychol. Sci.* **18**, 1077–1083 (2007).
- Ferguson, B. & Waxman, S. Linking language and categorization in infancy. *J. Child Lang.* **44**, 527–552 (2017).
- Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
- Mack, M. L., Love, B. C. & Preston, A. R. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proc. Natl. Acad. Sci. USA* **113**, 13203–13208 (2016).
- Theves, S., Neville, D. A., Fernández, G. & Doeller, C. F. Learning and representation of hierarchical concepts in hippocampus and prefrontal cortex. *J. Neurosci. J. Soc. Neurosci.* **41**, 7675–7686 (2021).
- Bowman, C. R., Iwashita, T. & Zeithamova, D. Tracking prototype and exemplar representations in the brain across learning. *eLife* **9**, e59360 (2020).
- Bowman, C. R. & Zeithamova, D. Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *J. Neurosci. J. Soc. Neurosci.* **38**, 2605–2614 (2018).
- Hirabayashi, T., Takeuchi, D., Tamura, K. & Miyashita, Y. Functional microcircuit recruited during retrieval of object association memory in monkey perirhinal cortex. *Neuron* **77**, 192–203 (2013).
- Lee, S. W., O'Doherty, J. P. & Shimojo, S. Neural computations mediating one-shot learning in the human brain. *PLoS Biol.* **13**, e1002137 (2015).
- Naya, Y., Sakai, K. & Miyashita, Y. Activity of primate inferotemporal neurons related to a sought target in pair-association task. *Proc. Natl. Acad. Sci. USA* **93**, 2664–2669 (1996).
- Naya, Y., Yoshida, M., Takeda, M., Fujimichi, R. & Miyashita, Y. Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. *Eur. J. Neurosci.* **18**, 2915–2918 (2003).
- Warren, D. E. & Duff, M. C. Not so fast: hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus* **24**, 920–933 (2014).
- Yanike, M., Wirth, S., Smith, A. C., Brown, E. N. & Suzuki, W. A. Comparison of associative learning-related signals in the macaque perirhinal cortex and hippocampus. *Cereb. Cortex* **19**, 1064–1078 (2009).
- van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J. & Fernández, G. Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *J. Neurosci. J. Soc. Neurosci.* **30**, 15888–15894 (2010).
- van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., Morris, R. G. M. & Fernández, G. Building on prior knowledge: schema-dependent encoding processes relate to academic performance. *J. Cogn. Neurosci.* **26**, 2250–2261 (2014).
- Walsh, C. R. & Rissman, J. Behavioral representational similarity analysis reveals how episodic learning is influenced by and reshapes semantic memory. *Nat. Commun.* **14**, 7548 (2023).
- Antony, J. W. et al. Semantic relatedness retroactively boosts memory and promotes memory interdependence across episodes. *eLife* **11**, e72519 (2022).
- Wang, Y., Mao, X., Li, B., Lu, B. & Guo, C. Semantic memory influences episodic retrieval by increased familiarity. *NeuroReport* **27**, 774–782 (2016).
- Wu, C. M., Meder, B. & Schulz, E. Unifying principles of generalization: past, present, and future. *Annu. Rev. Psychol.* <https://doi.org/10.1146/annurev-psych-021524-110810> (2024).
- Tenenbaum, J. B. & Griffiths, T. L. Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–640 (2001).
- Binder, J. R. et al. Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* **33**, 130–174 (2016).
- Hebart, M. N. et al. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife* **12**, e82580 (2023).
- Fang, Y. et al. Semantic representation in the white matter pathway. *PLoS Biol.* **16**, e2003993 (2018).
- Fernandino, L. et al. Concept representation reflects multimodal abstraction: a framework for embodied semantics. *Cereb. Cortex* **26**, 2018–2034 (2016).

30. Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
31. Barsalou, L. W., Dutriaux, L. & Scheepers, C. Moving beyond the distinction between concrete and abstract concepts. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **373**, 20170144 (2018).
32. Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A. & Rogers, T. T. Decoding semantic representations in mind and brain. *Trends Cogn. Sci.* **27**, 258–281 (2023).
33. Ghio, M. & Tettamanti, M. Chapter 52—Grounding sentence processing in the sensory-motor system. In *Neurobiology of Language* (eds Hickok, G. & Small, S. L.) 647–657 (Academic Press, San Diego, 2016).
34. Kiefer, M. & Pulvermüller, F. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex* **48**, 805–825 (2012).
35. Bi, Y., Wang, X. & Caramazza, A. Object domain and modality in the ventral visual pathway. *Trends Cogn. Sci.* **20**, 282–290 (2016).
36. Peelen, M. V. & Downing, P. E. Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* **93**, 603–608 (2005).
37. Martin, A. The representation of object concepts in the brain. *Annu. Rev. Psychol.* **58**, 25–45 (2007).
38. Caramazza, A. & Mahon, B. Z. The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends Cogn. Sci.* **7**, 354–361 (2003).
39. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. USA* **111**, 14565–14570 (2014).
40. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
41. McClelland, J. L. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J. Exp. Psychol. Gen.* **142**, 1190–1210 (2013).
42. Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends Cogn. Sci.* **15**, 527–536 (2011).
43. Ralph, M. A. L., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
44. Du, C. et al. Human-like object concept representations emerge naturally in multimodal large language models. *Nat. Mach. Intell.* <https://doi.org/10.1038/s42256-025-01049-z> (2025).
45. Xu, N. et al. Revealing emergent human-like conceptual representations from language prediction. *Proc. Natl. Acad. Sci. USA* **122**, e2512514122 (2025).
46. Xu, Q. et al. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nat. Hum. Behav.* **9**, 1871–1886 (2025).
47. OpenAI. Gpt-4o system card. Preprint at <https://arxiv.org/abs/2410.21276> (2024).
48. Bai, S. et al. Qwen2. 5-vl technical report. Preprint at <https://arxiv.org/abs/2502.13923> (2025).
49. Strachan, J. W. A. et al. Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-024-01882-z> (2024).
50. Griffiths, T. L. Bayesian Models of Cognition. In *Open Encyclopedia of Cognitive Science*. (eds Frank, M. C. & Majid, A.) (MIT Press, 2024). <https://doi.org/10.21428/e2759450.7b420317>
51. Bowers, J. S. & Davis, C. J. Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* **138**, 389–414 (2012).
52. Jones, M. & Love, B. C. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* **34**, 169–188 (2011).
53. Coutanche, M. N. & Thompson-Schill, S. L. Rapid consolidation of new knowledge in adulthood via fast mapping. *Trends Cogn. Sci.* **19**, 486–488 (2015).
54. Peelen, M. V. & Downing, P. E. Category selectivity in human visual cortex: beyond visual object recognition. *Neuropsychologia* **105**, 177–183 (2017).
55. van Kesteren, M. T. R. et al. Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia* **51**, 2352–2359 (2013).
56. O’Kane, G., Kensinger, E. A. & Corkin, S. Evidence for semantic learning in profound amnesia: an investigation with patient H.M. *Hippocampus* **14**, 417–425 (2004).
57. Postle, B. R. & Corkin, S. Impaired word-stem completion priming but intact perceptual identification priming with novel words: evidence from the amnesic patient H.M. *Neuropsychologia* **36**, 421–440 (1998).
58. Gabrieli, J. D., Cohen, N. J. & Corkin, S. The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain Cogn.* **7**, 157–177 (1988).
59. Naya, Y. Declarative association in the perirhinal cortex. *Neurosci. Res.* **113**, 12–18 (2016).
60. Quiroga, R. Q. Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
61. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
62. Cameron, K. A., Yashar, S., Wilson, C. L. & Fried, I. Human hippocampal neurons predict how well word pairs will be remembered. *Neuron* **30**, 289–298 (2001).
63. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* **3**, 946–953 (2000).
64. Fried, I., MacDonald, K. A. & Wilson, C. L. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 753–765 (1997).
65. McClelland, J. L. et al. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* **14**, 348–356 (2010).
66. De Brigard, F., Umanath, S. & Irish, M. Rethinking the distinction between episodic and semantic memory: insights from the past, present, and future. *Mem. Cogn.* **50**, 459–463 (2022).
67. Renoult, L. & Rugg, M. D. An historical perspective on Endel Tulving’s episodic-semantic distinction. *Neuropsychologia* **139**, 107366 (2020).
68. Renoult, L., Irish, M., Moscovitch, M. & Rugg, M. D. From knowing to remembering: the semantic–episodic distinction. *Trends Cogn. Sci.* **23**, 1041–1057 (2019).
69. Greenberg, D. L. & Verfaellie, M. Interdependence of episodic and semantic memory: evidence from neuropsychology. *J. Int. Neuropsychol. Soc.* **16**, 748–753 (2010).
70. Tulving, E. Episodic and semantic memory. In *Organization of Memory* (eds Tulving, E. & Donaldson, W.) 381–402 (Academic Press, 1972).
71. Irish, M. & Piolino, P. Impaired capacity for prospection in the dementias—theoretical and clinical implications. *Br. J. Clin. Psychol.* **55**, 49–68 (2016).
72. Duval, C. et al. What happens to personal identity when semantic knowledge degrades? A study of the self and autobiographical memory in semantic dementia. *Neuropsychologia* **50**, 254–265 (2012).
73. Irish, M., Addis, D. R., Hodges, J. R. & Piguet, O. Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. *Brain* **135**, 2178–2191 (2012).

74. Schacter, D. L. et al. The future of memory: remembering, imagining, and the brain. *Neuron* **76**, 677–694 (2012).
 75. Ferry, A. L., Hespos, S. J. & Waxman, S. R. Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Dev.* **81**, 472–479 (2010).
 76. Ferry, A. L., Hespos, S. J. & Waxman, S. R. Nonhuman primate vocalizations support categorization in very young human infants. *Proc. Natl. Acad. Sci. USA* **110**, 15231–15235 (2013).
 77. Perszyk, D. R. & Waxman, S. R. Infants' advances in speech perception shape their earliest links between language and cognition. *Sci. Rep.* **9**, 3293 (2019).
 78. Carey, S. The child as word learner. In *Linguistic theory and psychological reality* (eds. Halle, M., Bresnan, J. & Miller, G. A.) 264–293 (Cambridge, MA: MIT Press, 1978).
 79. Carey, S. & Bartlett, E. Acquiring a single new word. *Pap. Rep. Child Lang. Dev.* **15**, 17–29 (1978).
 80. Sun, H. et al. Brain-like functional organization within large language models. Preprint at <https://arxiv.org/abs/2410.19542> (2024).
 81. 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. J. Soc. Neurosci.* **34**, 163–170 (2014).
 82. Kriegeskorte, N. & Mur, M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 245 (2012).
 83. Yan, C. & Zang, Y. DPARSF: a MATLAB toolbox for 'pipeline' data analysis of resting-state fMRI. *Front. Syst. Neurosci.* **4**, 1377 (2010).
 84. Yan, C.-G., Wang, X.-D., Zuo, X.-N. & Zang, Y.-F. DPABI: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* **14**, 339–351 (2016).
 85. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).
 86. Ashburner, J. A fast diffeomorphic image registration algorithm. *NeuroImage* **38**, 95–113 (2007).
 87. Core Team. R. R: *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
 88. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multi-voxel pattern classification analyses. *NeuroImage* **59**, 2636–2643 (2012).
 89. Botta-Dukát, Z. Cautionary note on calculating standardized effect size (SES) in randomization test. *Community Ecol.* **19**, 77–83 (2018).
 90. Xia, M., Wang, J. & He, Y. BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS ONE* **8**, e68910 (2013).
- G. Zhang), and the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2022ZD0205104 to L. Zhu). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the paper.

Author contributions

Y.B. and G.Z. conceptualized and designed the research. G.Z. performed research and conducted data analysis. S.X. conducted the task in large language models. D.Z. and X.W. contribute valuable discussion to data analysis. L.Z. contributed valuable advice and discussion to both research conceptualization and data analysis. G.Z. and Y.B. wrote the original draft. L.Z., D.Z., and X.W. revised the manuscript. Y.B. supervised the research.

Competing interests

The authors declare no competing interests. All authors approved the final version of the paper for submission.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-72868-w>.

Correspondence and requests for materials should be addressed to Lusha Zhu or Yanchao Bi.

Peer review information *Nature Communications* thanks Marco Tettamanti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Acknowledgements

We thank Jian Li for helpful comments on earlier drafts of the manuscript. This work was supported by National Natural Science Foundation of China (NSFC, 32595491 to Y. Bi), the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2021ZD0204104 to Y. Bi), NSFC (32400869 to G. Zhang; T2421004, 32325023, 32441104 to L. Zhu; 32171052 to X. Wang), the China Postdoctoral Science Foundation (2023M740299, 2024T170062 to