

# Cortical knowledge structures guide word concept learning

Yanchao Bi

ybi@pku.edu.cn

State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

Guangyao Zhang

State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

Xiaosha Wang

Beijing Normal University

Dingchen Zhang

State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

Siwen Xie

Institute for Artificial Intelligence, Peking University

Lusha Zhu

Peking University <https://orcid.org/0000-0001-8717-6356>

---

## Biological Sciences - Article

### Keywords:

**Posted Date:** August 29th, 2025

**DOI:** <https://doi.org/10.21203/rs.3.rs-6982157/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** There is **NO** Competing Interest.

---

## Cortical knowledge structures guide word concept learning

Guangyao Zhang<sup>1</sup>, Xiaosha Wang<sup>1</sup>, Dingchen Zhang<sup>1</sup>, Siwen Xie<sup>4,6</sup>, Lusha Zhu<sup>2,3,5\*</sup> & Yanchao Bi<sup>2,3,4,5,1\*</sup>

1 State Key Laboratory of Cognitive Neuroscience and Learning & IDG/McGovern Institute for Brain Research, Faculty of Psychology, Beijing Normal University, Beijing, China.

2 School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

3 IDG/McGovern Institute for Brain Research, Peking University, Beijing, China

4 Institute for Artificial Intelligence, Peking University, Beijing, China

5 Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China

6 Yuanpei College, Peking University, Beijing, China

### Author Note

This work was supported by the STI2030-Major Project (2021ZD0204104 to Y. Bi, 2022ZD0205104 to L. Zhu), National Natural Science Foundation of China (Grant No. 32400869 to G. Zhang; Grant No.32171052 to X. Wang; Grant No.31925020, 82021004 to Y. Bi; Grant No. 32071095 to L. Zhu), the China Postdoctoral Science Foundation (Grant No. 2023M740299, 2024T170062 to G. Zhang), the Fundamental Research Funds for the Central Universities (to Y. Bi), and Center for Life Sciences at Peking University (to L. Zhu). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the paper.

\*Correspondence should be addressed to Yanchao Bi (E-mail address: ybi@pku.edu.cn) or Lusha Zhu (Email address: lushazhu@pku.edu.cn).

## Abstract

Human word-concept learning transcends simple associations between a word and referent exemplars, leveraging prior knowledge to generalize from few exemplars. Although Bayesian models explain such behavior, their neural underpinnings for prior structures and computations remain unclear. This study introduces a Neural Bayesian Model (NBM) to elucidate how prior knowledge representations guide new word learning. Using functional magnetic resonance imaging, we first measured the participants' neural activity during viewing familiar objects (and novel shapes as controls) to construct the neural prior space, and then the neural activity as participants learned new words associated with some of these visual stimuli. The NBM, which integrates neural representational priors derived from activities in ventral occipitotemporal cortex (VOTC), predicted new word neural representations and generalization behavior in learning with familiar objects, outperforming control models lacking neural priors. Conversely, hippocampal activity, not necessarily explained by the NBM, underpinned learning with novel shapes, reflecting a prior-free mechanism. Comparisons with large language models (LLMs) revealed LLMs' inferior alignment with human generalization, underscoring gaps in grounding word learning in nonverbal priors. These findings dissociate neural computational systems for concept learning: the VOTC mediates prior-based Bayesian inference, whereas the hippocampus supports exemplar-based associations. The results bridge computational theories of word learning with neural mechanisms, highlighting the dynamic interplay of semantic and episodic memory, and further promoting the incorporation of Bayesian-based learning mechanisms for LLM development.

## Introduction

What happens in the human brain when one learns that 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 exemplar 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 way by mapping words with corresponding exemplars, or a prior-based way with the help of our existing conceptual knowledge, particularly that regarding the categorical structures and relationships related to exemplars (Fig 1a)? It has been proposed that such background knowledge, likely grounded in our long-term semantic memory, acts to constrain candidate hypotheses about the word meaning and guide generalization, thereby enabling the rapid, data-efficient learning of word concepts (Xu & Tenenbaum, 2007). Such prior-based word learning is a powerful way for humans to construct and learn new word concepts (Ferguson & Waxman, 2017; Lupyan, 2006; Lupyan et al., 2007; Waxman & Markow, 1995) and is considered among the cornerstones of human intelligence, as well as an issue at stake in artificial intelligence models (Lake et al., 2015, 2017; Tenenbaum et al., 2011).

Despite its theoretical appeal, the neural mechanism by which the existing 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 concept learning research, derived from the associative learning literature, has examined the formation of artificially designed categorization rules or associations, during which the influence of prior knowledge is typically minimized. This line of studies converged on revealing the effects of the hippocampus and ventromedial prefrontal cortex (VMPFC) in learning new (conceptual) categories/associations (Bowman et al., 2020; Bowman & Zeithamova, 2018; Mack et al., 2016; Theves et al., 2021). In particular, the role of hippocampus and adjacent medial temporal regions, in learning via building new associations between stimuli (Hirabayashi et al., 2013; Lee et al., 2015; Naya et al., 1996, 2003; Warren & Duff, 2014; Yanike et al., 2009) has been highlighted, suggesting a prior-free learning mechanism in the hippocampus. Another line of studies specifically examined the effects of prior knowledge on learning novel associations and reported that the presence and strength of prior knowledge increased activity in the dorsomedial prefrontal cortex (DMPFC) and decreased the activity in the hippocampus (van Kesteren et al., 2010; van Kesteren et al., 2014). These findings resonate with the increasing attention paid to the interaction between semantic memory (i.e. long term store of the knowledge about the world) and episodic learning (Walsh & Rissman, 2023; Antony et al., 2022; Wang et al., 2016), but how exactly semantic memory neural representation participates in forming new word concepts remains unknown.

In comparison, 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 learning. A prominent class of models formalizes this process as Bayesian inferences over a structured hypothesis space of possible word meanings (Wu et al., 2024; Xu & Tenenbaum, 2007; Tenenbaum & Griffiths, 2001). Under this framework, learning involves evaluating each candidate hypothesis in the space, based on its initial plausibility (prior) and ability to explain the observed exemplars (likelihood). Generalization to new instances is then governed by integrating these possible hypotheses, and weighting each according to how strongly the learner believes in it after seeing the exemplars (posterior). Despite its simplicity, this probabilistic account of a putative hypothesis space captures key behavioral signatures of human word learning, including sensitivity to semantic structure and similarity among exemplars, graded generalization after a single example, and category-specific generalization with multiple examples. 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 shapes 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 how similar two objects are using a 7-point scale). While such a one-dimensional behavioral judgement can reflect aspects of prior knowledge on object relationships, it inevitably provides a flattened projection of the rich and high-dimensional organization of semantic knowledge (see discussions in Binder et al., 2016; Hebart et al., 2023). 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., 2018; Fernandino et al., 2016; Binder et al., 2009). 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, color, motion, and action), respecting an evolutionarily salient domain principle (e.g. the animate-inanimate principle reflected by the organized representations of animals, human faces and artifacts in human brain; Bi et al., 2016; Binder & Desai, 2011; Peelen & Downing, 2005; Martin, 2007; Caramazza & Mahon, 2003). Therefore, it is possible that the rich, structured, and distributed representations of object knowledge across the brain serves as a neurobiological base for prior beliefs central to Bayesian learning. However, which brain regions encode prior knowledge in word learning, and by what neural mechanisms this information guides word learning and generalization, remain elusive.

Here, we propose a Neural Bayesian Model (NBM), a neurobiologically grounded account of word concept learning in which 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 priors underlying Bayesian learning. Critically, it enables a systematic investigation – within a unified experimental and computational framework – of how word learning is shaped by different priors coded in brain regions previously implicated in distinct aspects of conceptual processing. In particular, the NBM allows us to examine how priors coded in the VOTC (implicated in object knowledge representation) and the hippocampus/VMPFC/DMPFC (implicated in learning novel associations and concepts) differentially support word learning and generalization, potentially reflecting their complementary roles in neurocomputation.

We applied the NBM to address five key questions. 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. spoon, eyeglass, hammer)? Second, are broader structured neural priors beyond specific exemplars necessary in 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, and can such a difference be explained by differential learning mechanisms? Fifth, does word concept learning in large language models (LLMs)—which do not explicitly implement the Bayesian model in training but show human-like language processing skills—compare with Bayesian word learning in predicting human behavior? Together, these questions position the proposed 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 and prediction computation; 2) Evaluating the effectiveness of the NBM in predicting the neural representations and behavior of new word learning, against multiple control models (prior-free model; permutation model; behavioral prior model) and conditions (weak prior stimuli – novel shapes) without structured neural prior considerations in both region of interests (ROI) and whole brain analyses; 3) comparing

Bayesian learning model with LLMs' fitness of human word concept learning behavior.

To obtain neural representational priors, we first measured brain activities of a set of objects covering the object domains that have been well studied ( $N = 58$ ) (fMRI Experiment 1). 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 of the NBM. A subset of these objects was then used as exemplars in a subsequent word learning experiment conducted inside scanner (fMRI Experiment 2; for an overview of study paradigms, see Fig 1b). In each trial, participants were first introduced to a novel word (e.g. 'Leca') together with several exemplars (e.g. spoon, eyeglass, hammer), and then asked to judge whether a probe object (e.g. axe) could also be described by the new word. These generalization judgments had no right or wrong answer, allowing us to measure subjective interpretations and their corresponding neural representations.

The key of this two-phase design is that we used the neural representations of each object obtained from Experiment 1, which was independent from learning, to construct the NBM and then to predict both the neural representations of new words and the generalization behavior obtained from Experiment 2 (i.e. the learning task). See Fig 1-2 for a schematic of the analysis pipeline and Methods for full methodological details, experimental procedure, and computational models. See Fig S1 for all the stimuli we used.

### **Behavior results**

Participants showed highly inter-subject behavioral response consistencies (ISCs) in both experiments (Experiment 1: oddball object similarity judgment task, Fisher-transformed ISCs = 0.517,  $P < .001$ ; Experiment 2: new concept learning generation to new objects task, Fisher-transformed ISCs = 0.483,  $P < .001$ ; see Section A in Supplementary Materials for details). Critically, in the word learning task (Experiment 2), the generation behavior replicated the classical result pattern (Xu & Tenenbaum, 2007), with sensitivity to the number and similarity of the referent exemplars in learning. When learning with a single exemplar, participants showed graded generalization to new objects that had high-, medium-, and low-level similarity with the learned exemplar; when learning with three exemplars, participants showed generalization sharpened into a much more all-or-none pattern depending on how similar the three exemplars were. We also replicated the predictive power of behavioral Bayesian model (BBM; Fig S2).

### **NBM**

#### ***Model overview***

The NBM formalizes word learning as a probabilistic inference over a structured neural hypothesis space (for flowchart of the NBM construction, see Fig 2). According to the model, learning a word such as 'Leca' involves evaluating a range of candidate hypotheses about what the word could refer to. Each hypothesis corresponds to a 'category' in semantic memory: 'Leca' might mean all spoons, tableware, or artifact items. These possibilities are structured hierarchically, forming an internal hypothesis space shaped by prior neural representations of object relationships.

Before seeing any new examples for learning, the learner assigns each hypothesis a prior probability, based on how likely that category is to be named. In line with previous work, 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-like hypothesis space. For instance, if 'tableware' form a well-separated cluster in the tree, it receives 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 labelled exemplars (e.g. spoon, eyeglass, hammer), they update their beliefs using Bayesian rule. The likelihood captures how well each hypothesis explains the observed examples, favoring smaller categories that tightly include the exemplars. Formally, the model assumes that the likelihood is inversely related to the size of the hypothesis (cluster size). The posterior then combines the prior and likelihood to yield an updated belief regarding the probable meaning.

Generalization involves integrating all plausible hypotheses. When deciding whether a new object (e.g. axe) could also be a 'Leca', the learner does not choose a single hypothesis with the highest posterior. Instead, they sum all hypotheses containing both axe and any exemplar in the internal hypothesis space.

### ***Constructing and validating neural hypothesis spaces***

The key idea of the NBM is to construct a prior hypothesis space directly from neural representations of long-term object knowledge and use this space to guide Bayesian word learning. We first obtained multivoxel response patterns for 58 objects across participants in Experiment 1, which is independent from word learning (Experiment 2). Within each ROI, we averaged the neural response elicited by each object across all the participants. The object-specific response patterns in 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 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 within 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 averaged activation map across all sample objects in Experiment 1 (Fig S3). As expected, a contrast analysis of well-established 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 face form area), animal pictures activated regions extending from the bilateral posterior fusiform cortex to lateral occipital cortex, and artifact pictures elicited activity in the bilateral lateral superior occipital cortex and other regions (Bi et al., 2016; Binder & Desai, 2011; Peelen & Downing, 2005; Martin, 2007; Caramazza & Mahon, 2003).

To assess the representational content captured by neural activation, and test whether it encodes richer information than behavioral judgements, we compared brain responses with behavioral similarity judgement along two important object properties – shape and semantic meaning – and low-level image properties. 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 rated the pairwise similarity of objects based on semantic and shape-based dimensions, yielding behavioral RDMs for each (Fig 1c). The two object property RDMs were significantly correlated ( $R = .668$ ), and exhibited lower and significant correlations with low level image pixel patterns. Using representational similarity analysis (RSA), we found that neural responses in different visual cortex regions tended to represent these two types of object properties to different degrees: the VOTC, defined by Harvard-Oxford Atlas, overall showed a unique effect of the behavioral-semantic RDM (partial  $Rho = .285$  with the other object property RDM and pixel-based-RDM controlled for,  $T = 9.369$ ,  $P_{\text{right-tailed}} < .001$ ); zooming into subsections of the VOTC, the neural RDMs of bilateral fusiform (defined by Harvard-Oxford Atlas) showed unique correlations with both behavioral-semantic (partial  $Rho = .231$ ,  $T = 7.460$ ,  $P_{\text{right-tailed}} < .001$ ) and - shape RDMs (partial  $Rho = .213$ ,  $T = 6.839$ ,  $P_{\text{right-tailed}} < .001$ ); 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 unique correlation with the behavioral-shape RDM of face pictures (partial  $Rho = .203$ ,  $T = 1.779$ ,

$P_{\text{right-tailed}} = .040$ ) and not artifact pictures (partial  $Rho = -.085$ ,  $T = -1.100$ ,  $P_{\text{right-tailed}} = .864$ ). By contrast, bilateral lateral occipitoparietal cortex (defined by Harvard-Oxford Atlas) showed a unique correlation with behavioral-shape RDMs (partial  $Rho = .292$ ,  $T = 9.608$ ,  $P_{\text{right-tailed}} < .001$ ). 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 values of neural priors. The neural RDMs are also shown in the hippocampus, VMPFC, and DMPFC (Fig S4), and no significant correlation was observed with either behavioral-semantic or -shape RDMs.

How does the neurally derived hypothesis space look, and to what extent is it similar to or different from the hypothesis space built from behavioral ratings? We compared the dendrogram computed based on the RDM in the VOTC with that derived from behavioral similarity judgements, following the same hierarchical clustering procedure. As shown in Fig S5, the two structures reveal broadly similar hierarchical structure. For instance, both trees clearly separate major object domains such as faces, animals, and artifacts. However, they also exhibit notable fine-grained differences. For example, while the spoon (object 19 in the VOTC-based dendrogram) and pen (object 20 in the VOTC-based dendrogram) are clustered closely together in the VOTC hypothesis space, 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 judgements, underscoring their value in revealing that internal knowledge representation likely informing word learning.

#### **Word concept learning with rich priors in the VOTC**

Having estimated this neural hypothesis space, we used it predict the neural representations of newly-learned words in Experiment 2 (Fig 2). For each word, the NBM would result in unique posterior probabilities in the neural hypothesis space, which represented the meaning of the word, and the generalization probability of each object in the corresponding domain belonging to a new word (e.g. 'Leca'). The neural representation of a new word was then predicted by averaging the prior neural representations of corresponding objects (obtained Experiment 1), weighted by their generalization probabilities.

We evaluated the effectiveness of the model in predicting neural representation and behavioral patterns via: 1) correlation between predicted and observed neural representations of words, and 2) correlation between predicted generalization probabilities to new probe objects and observed behavioral probabilities.

#### ***VOTC-based NBM predicts new word neural representations better than alternative models***

In the VOTC, the prediction of the NBM is highly significant in predicting the neural representations of new words (Pearson  $R_{\text{Fisher-z}} = 0.306$ ,  $BF_{10} = 9.630 \times 10^{10}$ ,  $T_{19} = 18.529$ ,  $P_{\text{right-tailed}} < .001$ ; Fig 3 and Table S2). That is, Bayesian model based on abstract structural properties of neural representations in the VOTC (height difference and cluster size) significantly predicted the neural encoding of meaning for a newly-learned word. When each of the three object domains was analyzed separately, the predictions were significant for all domains (Table S3). For simplicity, we report results collapsed across domains below. To evaluate the contribution of structured neural priors, we compared the NBM with a set of control models and learning condition: 1) model derived from a prior-free mechanism, 2) models constructed based on randomly permuted neural priors, 3) model constructed based on behavioral-rating priors, and 4) learning without rich priors.

#### **Does incorporating the neural priors matter? Comparing the NBM with the model without prior consideration**



To examine the extent to which priors contributes to word concept learning, we compared the NBM with a 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 exemplars spoon, eyeglass, and hammer in Experiment 2, the NMM predicted the new word neural representation as the mean of the neural patterns of these three objects (interpedently measured in Experiment 1). We found that the prediction of the NMM was significant (Pearson  $R_{\text{Fisher-z}}$  between predicted and observed neural activity pattern in the VOTC = 0.274,  $BF_{10} = 7.753 \times 10^{10}$ ,  $T_{19} = 18.294$ ,  $P_{\text{right-tailed}} < .001$ ). Critically, the predictive power of the NBM was significantly greater than that of the NMM (Pearson  $R_{\text{Fisher-z}} = 0.306$  v.s. 0.274,  $BF_{10} = 7.934 \times 10^9$ ,  $T_{19} = 16.653$ ,  $P_{\text{two-tailed}} < .001$ ; Fig 3 and Table S2). In addition, after controlling the predicted neural pattern of the NMM using partial correlation, the NBM still had unique predicative power (Pearson  $R_{\text{Fisher-z}} = 0.126$ ,  $BF_{10} = 1.011 \times 10^{11}$ ,  $T_{19} = 19.353$ ,  $P_{\text{right-tailed}} < .001$ ; Fig 3 and Table S2). These results indicated that incorporating broader, structured neural priors indeed contributed to the predictive power of the NBM.

#### Does incorporating the specific neural-priors matter?

To test whether the advantage of the NBM over the NMM 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 permutations 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 (Observed Pearson  $R_{\text{Fisher-z}} = 0.306$ , mean of null distribution = 0.304, Standard Effect Size (SES, calculated for each effect as the difference between the observed value and the mean value of the null distribution, divided by the standard deviation of the null distribution) = 7.569,  $P_{\text{right-tailed}} < .001$ ; Fig 3 and Table S2), indicating that specific structured neural priors matter for the predictive power of the NBM.

#### Does incorporating neural priors have predictive power beyond behavioral priors?

Neural priors have the benefit of respecting the distributed, high-dimensional representation properties of priors, as opposed to behavior-rating 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, by incorporating such neural priors, we compared it with the BBM, which was constructed using the same procedure as the NBM, except that the prior computation was based on the behavioral ratings of the semantics distance of sample objects. The results showed that after controlling the predicted pattern from the BBM using partial correlation, the NBM had a unique predictive power for the learned word neural representation in the VOTC (Person  $R_{\text{Fisher-z}} = 0.042$ ,  $BF_{10} = 4.851 \times 10^6$ ,  $T_{19} = 10.043$ ,  $P_{\text{right-tailed}} < .001$ ; Fig 3 and Table S2).

#### Does the NBM specifically apply to learning with rich priors? Comparing word concept learning with familiar objects and novel shapes

Another way to test the advantage of the NBM driven by the inclusion neural prior 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 representation). The results showed that although the prediction accuracy of the NBM in learning with novel shapes was significant (Pearson  $R_{\text{Fisher-z}} = 0.284$ ,  $BF_{10} = 5.735 \times 10^{11}$ ,  $T_{19} = 20.568$ ,  $P_{\text{right-tailed}} < .001$ ), the NBM in

learning with familiar objects outperformed the NBM in learning with novel shapes (Pearson  $R_{\text{Fisher-z}} = 0.306$  v.s. 0.284,  $BF_{10} = 4.606$ ,  $T_{19} = 2.802$ ,  $P_{\text{two-tailed}} = .011$ ; Fig 3 and Table S2). That is, the NBM exhibits greater power in predicting learning with richer priors.

### ***VOTC-based NBM predicts generalization behavior***

In the second analysis, we evaluated the model performance by computing the Pearson correlation between the NBM-predicted probability of generalization behavior and the actual behavioral responses (the percentage of participants judging yes to the probe object belonging to the learned word). The results showed that the NBM significantly predicts generalization behavior (Pearson  $R = .288$ ,  $BF_{10} = 21.466$ ,  $T_{94} = 2.915$ ,  $P_{\text{right-tailed}} = .002$ ; Fig 4 and Table S2). Similar to the neural analyses above, we compared the predicative power of the NBM to the same control models and learning condition, revealing consistent results.

### ***The NBM outperforms alternative models in predicting generalization behavior.***

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 predicting generalization behavior during word concept learning. In the NMM, the generalization behavior to a probe object is predicted by the similarity of the probe object's neural response to the NMM-predicted word concept representation, that is, quantified as the Pearson  $R$  between the neural pattern of the probe objects (obtained in Experiment 1) with the NMM-predicted neural pattern of the new word. Correlating these predicted values with the observed human behavior revealed that the NMM failed to make effective prediction (Pearson  $R = -.044$ ,  $BF_{10} = 0.173$ ,  $T_{94} = -0.429$ ,  $P_{\text{right-tailed}} = .666$ ; Fig 4 and Table S2). A direct magnitude comparison showed the advantage of the NBM over the NMM ( $z = 2.927$ ,  $P_{\text{two-tailed}} = .003$ ; Fig 4 and Table S2). After controlling for the predicted values of the NMM using partial correlation, the NBM still made significant predictions (Pearson  $R = .320$ ,  $BF_{10} = 54.923$ ,  $T_{94} = 3.270$ ,  $P_{\text{right-tailed}} = .002$ ; Fig 4 and Table S2). Examining the power of the NBM in predicting behavior relative to prior-permutation-control models (the same control models as in the section above but with 1000 permutation iterations without group-level bootstrap) again revealed the significant effect of the NBM (mean of null distribution = .014,  $SES = 2.048$ ,  $P_{\text{right-tailed}} = .020$ ; Fig 4 and Table S2).

### ***The NBM has additional contributions to the BBM in predicting generalization behavior.***

First, replicating previous literature (Xu & Tenenbaum, 2007), the BBM significantly predicted generalization behavior (Pearson  $R = .907$ ,  $BF_{10} = 5.779 \times 10^{32}$ ,  $T_{94} = 20.847$ ,  $P_{\text{right-tailed}} < .001$ , Fig S2). Critically, after controlling the predicted values of the BBM using partial correlation, the NBM still had uniquely significant effects in predicting generalization behavior (Pearson  $R = .192$ ,  $BF_{10} = 2.372$ ,  $T_{94} = 1.894$ ,  $P_{\text{right-tailed}} = .031$ ; Fig 4 and Table S2), indicating the specific contributions of neural priors in predicting generalization behavior.

### ***The NBM specifically applies to word concept learning with rich priors.***

We found that the NBM could only predict generalization behavior in learning with familiar objects (results reported above) not novel shapes (Pearson  $R = .277$ ,  $BF_{10} = 1.557$ ,  $T_{22} = 1.350$ ,  $P_{\text{right-tailed}} = .095$ ; Fig 4), although the difference between these two learning conditions was not significant ( $z = 0.049$ ).

### ***Summary for the VOTC***

In the VOTC, a region known for storing rich object knowledge, the NBM significantly predicted both the neural representation of newly learned word concepts and participants' generalization behavior in word learning with familiar objects (with rich priors). Such a Bayesian model with neural priors 1) outperforms neural mean-

models (i.e. neural models without broader structured prior considerations) and prior-permutation-control models, 2) has additional predictive power beyond using behavioral priors in the Bayesian model, and 3) has more advantages in explaining word learning with familiar objects than with novel shapes (i.e. weaker priors). These results showed the adequacy and necessity that the brain regions representing prior knowledge are involved in word concept learning in a manner consistent with Bayesian inference mechanisms.

#### **Word concept learning with rich priors in the hippocampus and VMPFC/DMPFC**

##### ***Hippocampus- VMPFC- and DMPFC-based NBM failed to predicting new word neural representations or generalization behavior***

Surprisingly, the same NBM, when constructed using dendrograms derived from neural activity in the hippocampus, VMPFC, and DMPFC, failed to predict the neural representation of newly learned word concepts represented in these respective regions (for the hippocampus, Pearson  $R_{\text{Fisher-z}} = 0.006$ ,  $BF_{10} = 0.512$ ,  $T_{19} = 0.865$ ,  $P_{\text{right-tailed}} = .199$ ; for the VMPFC and DMPFC,  $P_{\text{right-tailed}} > .5$ ) or participants' generalization behavior from familiar objects ( $P_{\text{right-tailed}} > 0.300$ ; Fig 5a and Table S2).

We further considered the possibilities that these regions do play a role in Bayesian word learning, but in ways different from those observed in the VOTC: 1) these regions may not encode structured priors themselves but instead rely on prior knowledge represented in cortical regions such as the VOTC; 2) they may be involved in tracking the dynamic updating of conceptual representations over the course of learning (see Theves et al. 2021); and 3) they may contribute selectively to learning with weak priors.

##### **The NBM constructed based on neural priors in the VOTC.**

Is it possible that these regions participate in forming the new word neural representation, but rely on prior knowledge stored in other brain areas, such as the VOTC? To test this hypothesis, we performed RSA to compare RDM of the posterior probability patterns in the NBM constructed in the VOTC and the RDM of the observed neural patterns of new words in the hippocampus, VMPFC and DMPFC (Fig 5b). 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 (for the VMPFC, Spearman  $Rho_{\text{Fisher-z}} = 0.030$ ,  $BF_{10} = 0.505$ ,  $T_{19} = 0.854$ ,  $P_{\text{right-tailed}} = .202$ ; for the hippocampus and DMPFC,  $P_{\text{right-tailed}} > .9$ ; Fig 5b and Table S4). The VOTC-based NBM-model-RDM did significantly correlate with neural RDMs in the VOTC (Spearman  $Rho_{\text{Fisher-z}} = 0.156$ ,  $BF_{10} = 92.148$ ,  $T_{19} = 4.000$ ,  $P_{\text{right-tailed}} < .001$ ; Fig 5b and Table S4), replicating the results in the above sections. That is, even considering prior building from the VOTC, neural activity patterns of the hippocampus/VMPFC/DMPFC during new word learning was not effectively predicted by the Bayesian model.

##### **Concept representation updating.**

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 learning, such as tracking and updating the internal representation of concept as learning unfolds (Theves et al., 2021)? That is, as the new exemplar comes in, the neural representation updates accordingly, reflected by the changes in the posterior probability pattern across the hypotheses in the NBM, these regions, as regions sensitive to learning, might be sensitive to the magnitude of such changes. We conducted a parametric modulation analysis to test this possibility (Fig 5c), examining whether the activation strength for the first word learning event (exemplar 1), the second word learning event (exemplar 2), and the third word learning event (exemplar 3), was a function to how much the posterior pattern across hypotheses changes. We constructed the VOTC-based NBM to compute the updating for each of the learning

stage (Fig 5c): For exemplar 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 between the prior probability pattern with its posterior probability pattern. The parametric modulation results revealed that the VOTC-based NBM did not predict the activation strength in the hippocampus, VMPFC or DMPFC ( $P_{\text{right-tailed}} > .4$ , Fig 5c and Table S5), but did in the VOTC (Beta = 1.794,  $BF_{10} = 2.275$ ,  $T_{19} = 1.987$ ,  $P_{\text{right-tailed}} = .031$ ; Fig 5c and Table S5). Thus, there was no evidence that the hippocampus/VMPFC/DMPFC keeps track of the neural pattern changes predicted by the Bayesian model.

#### Predicating word concept learning with weak priors.

Interestingly, while the above results showed that the NBM in the hippocampus/VMPFC/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  $R_{\text{Fisher-z}} = 0.020$ ,  $BF_{10} = 4.399$ ,  $T_{19} = 2.389$ ,  $P_{\text{right-tailed}} = .014$ ; Fig 5d and Table S6). This was not observed in the VMPFC or DMPFC ( $P_{\text{right-tailed}} > .7$ ). Notably, the NMM also showed significant predictions in the hippocampus (NMM: Pearson  $R_{\text{Fisher-z}} = 0.021$ ,  $BF_{10} = 6.970$ ,  $T_{19} = 2.652$ ,  $P_{\text{right-tailed}} = .008$ ), and there was no significant difference in the predictive power between the NMM and the NBM (difference = 0.001,  $BF_{10} = 0.238$ ,  $T_{19} = -0.236$ ,  $P_{\text{two-tailed}} = .816$ ; Fig 5d and Table S6). Partial correlation analysis revealed no additional power for either the NBM beyond the NMM ( $P = .065$ ) or the NMM beyond the NBM ( $P = .386$ ) in the hippocampus (Table S6).

#### **Summary**

For the hippocampus/VMPFC/DMPFC, we did not observe evidence for their participation in learning with rich priors: the NBM failed to predict their neural representations of the new words; using neural priors from the VOTC also yielded null results; there was no evidence that they are sensitive to the new word concept updating process during learning. By contrast, when looking into word learning with weak priors, the NBM did predict the new word representations in the hippocampus, yet with comparable predictive power to the NMM, with both models showing stronger effect in predicting the new word neural representation of word tokens based on novel shapes than that based on familiar objects (Table S6). That is, hippocampal neural representation of new words is a function of the associated novel exemplars, without prior contributions.

#### **Double dissociation between word concept learning with rich and with weak priors**

In the above ROI analyses, we observed different patterns regarding the effects of the NBM in learning with rich or weak prior, across ROIs: in the VOTC, which has been implicated in object representation, the NBM has greater power in learning with familiar objects than in learning with novel shape; in the hippocampus, which has been previously implicated in learning concepts from novel and/or artificially designed stimuli, only neural representations of the new words learned with novel shapes, and not with familiar objects, are predicted by the NBM. These findings highlight the commentary roles of these regions in Bayesian word learning. To examine this dissociation pattern more directly, we performed a two-way interaction analysis (brain region x concept type: VOTC/hippocampus  $\times$  familiar objects/novel shapes). A significant two-way interaction effect was observed ( $F_{(1,19)} = 10.304$ ,  $P_{\text{two-tailed}} = .005$ ; Fig 5e). Further simple effect analyses revealed that the NBM had better predictive power in learning with familiar objects than with novel shapes in the VOTC (difference = 0.021,  $BF_{10} = 4.606$ ,  $T_{19} = 2.802$ ,  $P_{\text{two-tailed}} = .011$ ), and better predictive power in learning with novel shapes than with familiar objects in the hippocampus (difference = 0.015,  $BF_{10} = 1.794$ ,  $T_{19} = -2.254$ ,  $P_{\text{two-tailed}} = .036$ ). This suggested a

double dissociation between the hippocampus and the VOTC in concept learning: the VOTC is involved in learning with rich priors, whereas the hippocampus is involved in learning with weak priors.

#### **Whole-brain searchlight results in predicting new word neural representations (with rich prior)**

The whole-brain searchlight analysis employed the same methods as the ROI analysis and revealed similar result to those in the VOTC. 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  $P_{\text{right-tailed}} < .001$  using cluster-level FWE correction  $P < .05$ , Fig S6 and Table S7), which were consistently found to represent object knowledge (Bi, et al., 2016; Binder & Desai, 2011; Caramazza & Mahon, 2003; Martin, 2007; Lambon Ralph et al., 2017). The most robust effects were obtained in bilateral occipitotemporal fusiform gyrus (within the VOTC), where the NBM showed greater predictive power than both the NMM and prior-permuted control models, additional predictive power beyond the BBM, and greater predictive power in learning with familiar objects than with novel shapes (Fig S6 and Table S7-S8). No results were found in the hippocampus/VMPFC/DMPFC. These results, together with the ROI results, converge to show that the brain regions representing object knowledge support word concept learning by following Bayesian inference mechanisms. For more detailed information, please see the Supplementary Materials.

#### **Bayesian model predicts human concept learning behavior better than LLM**

The recent generation of LLMs, while do not explicitly implement the Bayesian model in training, shows human-like language processing skills. Thus, we considered this new type of language-processing models as another type of control, with which the Bayesian model of word learning is compared. We tested three state-of-art LLMs that can process multimodal inputs: GPT-4o-2024-11-20 (Open AI, 2024), Qwen2.5-VL (Bai et al., 2025). We prompted these models to perform the same task as in Experiment 2 (learning experiment), conducted in independent chat sessions over 20 repetitions per word-concept (following a procedure similar to Strachan et al., 2024), setting temperature=1.0 and top-p=1.0 to reflect the original distribution predicted by the models (see Fig 6 for details). The 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 showed that GPT-4o and Qwen2.5-VL significantly correlated with human word concept learning generalization behavior when combining learning with both familiar objects and novel shapes (Table 1 and Fig 6).

Since the VOTC-based NBM had additional contributions to the BBM in predicting generalization behavior, here we calculated the best predictions of Bayesian learning model by combining the NBM and the BBM, and then compared it to the LLMs. To obtain the best prediction, we fit the observed generalization behavior against the predictions of both the NBM and the BBM via a general linear model, where the generalization behavior was predicted by considering the contributions from both the NBM and the BBM. The Bayesian learning model correlated with human behavior very well ( $R = .903$ ,  $P < .001$ ) and significantly more strongly than the LLMs ( $z_s > 5.833$ ,  $P_s < .001$ ). These advantages persisted when analyzing learning with familiar objects only ( $z_s > 6.711$ ,  $P_s < .001$ ). In learning with novel shapes, this advantage was absent between the Bayesian learning model ( $R = .903$ ,  $P < .001$ ) and GPT ( $R = .795$ ,  $P < .001$ ; difference,  $z = 1.628$ ,  $P = .104$ ; Fig 6 and Table 1), yielding a significant two-way interaction (model  $\times$  concept type: Bayesian learning model/GPT  $\times$  familiar objects/novel shapes;  $z = 3.738$ ,  $P < .001$ ). That is, despite the billions of parameters ( $\geq 7B$ ) and the super-rich prior knowledge in these LLMs obtained in pre-training (Bai et al., 2025; Open AI, 2024), these models do not simulate human new word learning behavior as well as Bayesian learning model, especially when learning with rich priors.

## Discussion

To understand how the human brain learns new word concepts based on prior knowledge, we developed and tested the NBM. This model is derived from the hypothesis that learning new words considers not only the representations of exemplars directly associated with, but also the broader background prior neural structure, as a Bayesian inference process. The results showed the advantages of the NBM over control models without incorporating such structured neural prior and over items without rich priors, in predicting both neural representations in the VOTC and generalization behavior in learning with familiar objects. Meanwhile, neural activity in the hippocampus did not show effects on learning with familiar objects but with novel shapes (weak priors). We discuss these key findings in turn below.

### **VOTC: Supporting prior-based new word neural representation and behavior by Bayesian neural computations**

In learning with familiar objects, the NBM successfully predicted new word neural representation in the VOTC, along with other regions commonly observed to represent object knowledge. The NBM considers the prior representational structure of not only associated exemplars when forming the new word representation. It not only better predicts the neural representation of the learned word, but also significantly predicts the subsequent generalization behavior – how likely 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 (Griffiths, 2024; Bowers & Davis, 2012; Jones & Love, 2011), indeed offers a computational framework for neural mechanisms that support learning with existing representations in semantic memory. This is in line with a hypothesis that word concept learning may happen based on a ‘fast mapping’ mechanism, that such learning integrate information into memory networks of neocortex (Coutanche & Thompson-Schill, 2015), especially when the new information is consistent with prior information (Kumaran et al., 2016; McClelland, 2013). New word concepts update the existing neural representation in the long-term (semantic) memory store directly (Kumaran et al., 2016; Coutanche & Thompson-Schill, 2015; McClelland, 2013). Note that for neural representation here, even for novel objects, the VOTC-based NBM had significant predictions. We contend that, even for a novel shape without clear prior explicit semantic knowledge, there is a certain perceptual priors regarding the shape space that contribute to forming the shape categories. We do not necessarily ascribe the entailed representation (learned categorical/concept representations and prior ones) to be object semantics – it might be the shape or whatever contents of stored information contents here.

### **Hippocampus: Supporting new word neural representation for novel stimuli associations**

When learning the association between a word token and visual stimuli without rich prior (novel shapes), the learned neural representation in the hippocampus is significantly explained by the composite of the neural representations of only the associated exemplar themselves (i.e. simple means), with additional consideration of the prior structure adding no explanatory power (i.e. the NBM not superior to the NMM). This result is in line with the literature on associative learning, showing that the hippocampal response (fMRI response or new neuronal response modulation) is sensitive to learning specific associations of objects (Hirabayashi et al., 2013; Lee et al., 2015; van Kesteren et al., 2013, 2014; Warren & Duff, 2014), or specific combinations of attributes (Bowman et al., 2020; Bowman & Zeithamova, 2018; Mack et al., 2016), which are inherently more episodic in nature.

What was not predicted was the absence of effects in the hippocampus in learning with familiar objects. It did not result in the new word neural representations that can be predicted by the NBM. Even using neural priors in the VOTC failed to make predictions in the hippocampus. Again this observation is consistent with the postulation of the ‘fast mapping’ mechanism, that such learning information consistent with priors can bypass the hippocampus and gets integrated into the cortical semantic memory stores (Kumaran et al., 2016; McClelland,

2013). These findings here corroborate the hypothesis that learning word concepts with rich prior may bypass the hippocampus, or at least involve the hippocampus in a way that is different from learning novel information. The key point is that both the NBM and NMM have a significantly stronger effect on predicting the new word neural representation in learning with novel shapes, than that with familiar objects. How can we reconcile the classical evidence that hippocampal lesions lead to semantic learning deficits in cases such as H.M. (O’Kane et al., 2004; Postle & Corkin, 1998; Gabrieli et al., 1988)? We speculate that the deficits arise from episodic aspects of word learning, such as the need to remember the specific word forms, which are not tested in the current context.

#### **Negative findings in the medial prefrontal, and medial temporal regions**

We did not observe significant effects of either model on learning word with familiar objects or novel shapes in the medial prefrontal lobe, in predicting either learned neural representation or concept representation updating as a function of changes in the VOTC-representations during learning. These null results contrasts with positive findings regarding the involvement of the medial frontal cortex in learning categories or associations (Theves et al., 2021; Bowman et al., 2020; Bowman & Zeithamova, 2018; van Kesteren et al., 2014, 2013, 2010). In addition, medial temporal regions beyond the hippocampus including the perirhinal cortex that have been shown to encode specific learned associations or concepts (e.g. Naya, 2016; Hirabayashi et al., 2013; Quiroga, 2012; Quiroga et al., 2005; Naya et al., 2003; Cameron et al., 2001; Kreiman et al., 2000; Fried et al., 1997; Naya et al., 1996), yet did not exhibit significant effects in our whole-brain analyses. The negative findings were difficult to interpret. One possibility is that they still engage in word concept learning and establish relationships with associated stimulus representations, but not in the ways we tested here, or that they are sensitive to associative representations related to a context broader than the limited object space sampled in our experiment. The exact computational manner of these regions in word-concept learning remains to be understood.

#### **Broader implications for Bayesian, semantic, and episodic learning**

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 realized in the brain (McClelland et al., 2010). Here, the observation that neural representations in the VOTC, but not the hippocampus, predicted both the neural representation of concepts 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 learning may be instantiated in the representational topology of the cortical systems, and word concept learning may proceed through reshaping this topology. More broadly, rational models such as Bayesian inferences are sometimes criticized as overidealizing human cognition, despite evidence that human behavior is frequently biased and variable. By grounding priors and inferences of concepts into the distributed, graded, and individually heterogeneous neural codes, our approach opens a path toward understanding how individual differences in word learning and generalization emerge from variability in the neural priors.

This framework also have implications for the intricate relationship between episodic and semantic memory, which has been long recognized (De Brigard et al., 2022; Renoult & Rugg, 2020; Renoult et al., 2019; Greenberg & Verfaellie, 2010; Tulving, 1972). Recent evidences showed that these two types of memory share large overlapping neural correlates (as reviewed by Renoult et al., 2019), the relationships between stimuli in semantic memory can influence their performance in episodic memory tasks (Walsh & Rissman, 2023; Antony et al., 2022; Wang et al., 2016), and that impairments in one can affect the function of the other (Irish & Piolino, 2016; Duval et al., 2012; Irish et al., 2012; Schacter et al., 2012). Our findings add to this line of insights about the dynamic

interplay of episodic and semantic memory – learning associations, a classical episodic learning task, in the context of 1) associating a word token with familiar objects and 2) with an instruction and task requirement that encourages categorization and generalization, is supported by Bayesian inference computation of prior (semantic) long-term neural representations in the cortex, which is different from the results reported about learning arbitrary associations of specific stimuli implicating the hippocampus. It is important to note that we do not know if these two factors (word-like token; context/tasks) are necessary for these effects to occur. Developmental research has highlighted the specific roles of verbal cues, reporting that children only use new word cues with phonological patterns of their native language to learn the category based on associated objects (and to generalize accordingly), and not with other cues, such as sounds of different linguistic groups, scrambled speech, or animal sounds (Ferry et al., 2010, 2013; Perszyk & Waxman, 2019). 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; Carey, 1978; Carey & Bartlett, 1978). Whether Bayesian-inference-computation neural mechanisms explain these behavioral observations during development is an important questions for future research.

The target neural model was inspired by and developed based on, Bayesian inference model for human word learning in the real world, that is, linking words with external referents and nonverbal experiences. The recent generation of LLMs, without implementing such Bayesian mechanisms in training, has nonetheless shown unprecedented success in processing language in a human-like manner, capturing internal representations with structures similar to those of human concepts, measured both behaviorally and neurally (Sun et al., 2024; Du et al., 2025). However, when learning new words by linking them with pictures, the two state-of-art LLMs that we tested -- GPT-4o and Qwen2.5-VL – captured human word learning behavior much worse than the Bayesian models. This disadvantage was absent in GPT-4o and the Bayesian model when learning with novel shapes. Several alternative interpretations warrant further consideration. One possibility is that learning with familiar objects (rich priors) requires stronger Bayesian inference mechanisms than learning with novel shapes (weak priors), which are not explicitly implemented when training LLMs. An other possibility is that processing the visual shape information of novel shapes recruits similar underlying mechanisms between GPT-4o and humans without extra contributions from prior knowledge. Filling such gaps in the training/learning mechanism holds promise for the development of language models that improve grounding word learning in nonverbal experiences.

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



## Methods

### Ethic 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 the informed consent form before taking part 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 and native Chinese speakers. None of them had experienced psychiatric or neurological disorders or had sustained a head injury. The sample size was 10 (8 women; mean age, 21.20 years; s.d., 1.14 years) for the pilot experiment, and 20 (13 women; mean age, 21.45 years; s.d., 2.06 years) for Experiment 1 and 2 and the semantic distance-judgement task. Each participant in pilot experiment received payments of 30 RMB. Each participant in Experiment 1 and 2 received payments of 230 RMB. Another 20 participants (12 women; mean age, 21.80 years; s.d., 1.99 years) participated in the visual-shape distance-judgement task of familiar objects with each participant receiving payments of 30 RMB.

### Designs and procedures

The experimental materials contained 58 gray object pictures of 300\*300 pixels, with 13 animals, 13 famous human faces, 19 human-made artifacts and 13 novel shapes. In each domain, 8 objects were used as probe objects and the rest ones were used as exemplars during learning. All stimuli were shown in Fig S1. The pilot experiment had the same experimental procedure as Experiment 2 but outside the fMRI scanner, to ensure the feasibility of the procedure within fMRI scanning session.

In Experiment 1, participants were asked to view each of the object pictures and, think about their meaning, while undergoing fMRI scanning. Each picture was presented for 1s, followed by one fixation interval. Each fixation lasted for 1-3s, with an average duration of 2s. When the fixation point was red, participants needed to judge whether the object after this fixation point was similar to the last object, and then press a key to make their judgment. This task included 10 runs of 4min and 28s each, and employed an event-related design. Each run contained 58 object pictures, and 11 object pictures as catch trials. Fifteen new words were also included as filler trials. The numbers and orders of trials for different stimuli were counterbalanced across runs and participants.

In Experiment 2, participants needed to perform the new word learning task in which they were asked to help a friend who spoke another language to choose the object she or he wanted. In each trial, this friend would show either one or three exemplars in turns, each labelled with one word token lasting for 2s. Subsequently, the word token appeared singly again for 3s, followed by a probe object lasting for 2s. The exemplars, single new word and new probe object were separated by one fixation interval. Each fixation appeared for 0.5-1.5s, with an average duration of 1s. Participants needed to learn the new word's concept based on the exemplars and then press a key to judge whether the new probe object was the one their friend wanted. We manipulated the number and similarity of exemplars, leading to 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 contains eight probe objects which would not be used as exemplars (Fig S1). There were also three other artifact words that differentiate between visual and functional similarities (exemplars that were highly similar in only visual aspects, only functional aspects, both aspects), leading to 15 different word concepts in total (12 ones of familiar-object exemplars and three ones of novel-shape exemplars). This task included

4 runs of 8min and 28s each. Each run started and ended with a fixation lasting 10s. Each word token had two trials in each run, with the number and order of trials for different words counterbalanced across runs and participants.

After fMRI scanning, participants in Experiment 1 and 2 were asked to perform an online semantic distance-judgement task using a multi-arrangement paradigm (Kriegeskorte & Mur, 2012) via NAODAO (<https://www.naodao.com/>). In this paradigm, participants dragged and dropped the 58 objects in 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 for 30 minutes. The visual-shape distance-judgement task shared the same procedure except that participants were asked to arrange the 45 familiar objects according to their visual-shape distance.

Finally, we asked two LLMs, including GPT-4o-2024-11-20 and Qwen2.5-VL-7B-Instruct to perform the same word learning task in Experiment2. For a given trial, supposing deciding whether the axe belongs to 'Leca' given the exemplars of the spoon, hammer and eyeglass, the exemplars were presented in three separate messages, mirroring the setting in human experiments (see Fig 6 for prompts). Each trial was repeated 20 times in an independent chat session, with temperature=1.0 and top-p=1.0 to reflect the original distribution predicted by the models. The generalization probability for a given trial was averaged across the 20 iterations.

### **Image acquisition and preprocessing**

All functional and structural MRI data were collected on a 3T Siemens Trio Tim 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 V4.3 (Yan & Zang, 2010) implemented in DPABI V3.0 (Yan et al., 2016). 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 module (Ashburner & Friston, 2005). Next, a custom, study-specific template was generated by applying diffeomorphic anatomical registration through exponentiated lie algebra (DARTEL; Ashburner, 2007). 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 multi-voxel pattern analysis (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 customized R script (R Core Team, 2020) unless specifically stated.

## First-level analysis

At the first level, a general-linear-model (GLM) analysis was performed to explore the fixed effect of each regressor for each participant in Experiment 1 and 2, respectively. In Experiment 1, each of the 58 objects were 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 interests were modelled, one for the exemplars and the other for the single new word. All probe objects were modeled as one nuisance regressor. In each experiment, six head motion parameters obtained by head motion correction were also included as nuisance regressors, and a high-pass filter (128 seconds) was used to remove low-frequency signal drift for each run. The group-level averaged t-value map across the participants for each of the 58 objects in Experiment 1 was estimated after GLM analyses, which was used as the neural priors. The t-value map of the exemplars together with the single new word for each word and participant was estimated in Experiment 2 after GLM analysis, which was used as the new word neural representation.

The group-level t-value map for each of the 45 familiar objects were first utilized to validate the advantage of neural priors, i.e. respecting the diverse types of representations of objects distributed in the brain. For each brain region, the activation pattern of each familiar object was extracted, according to which the dissimilarity (1-R) of each pair of objects was calculated, resulting in the neural RDM. Meanwhile, the behavior-rating semantic and shape-based RDM for each pair of familiar objects were estimated according to the distance-judgement tasks, resulting in two targeted model-RDMs. Another pixel-based RDM was also considered. For each region, partial correlation analyses were conducted to calculate the unique correlation (spearman-Rho) of the neural RDM to each of targeted model-RDM, with the other targeted model-RDM and the pixel-based RDM controlled.

Model construction, prediction and comparison analysis were conducted via both the whole-brain searchlight MVPA (with 5 x 5 x 5 cube size centered on each voxel) and ROI-based MVPA. The whole-brain searchlight analysis was conducted within the group-based gray mask. To obtain the mask, the normalized structural image was segmented into different tissues for each participant. The resulting gray matter probabilistic images were resliced to the same spatial resolution as that of the functional image, averaged across participants, and thresholded at 0.25 to generate a binary mask for searchlight mapping. All prediction and comparison procedures at both the whole-brain and ROI levels were implemented by a custom script in R.

## Model construction and comparison

### NBM

For a given region, the NBM was constructed based on the group-averaged t-value patterns of each of the 58 objects observed in Experiment 1. A hierarchical clustering analysis was performed based on the dissimilarities in t-value patterns among the 58 objects to construct the prior knowledge dendrogram using *complete linkage* (defaulted in R) to calculate the maximum distance between clusters before merging. Each node in the dendrogram represented a possible hypothesis. The prior, likelihood and posterior of each node and the generalization behavior of the probe object were calculated as follows:

*Prior probability computation.* Following the previous literature, the prior probability of each node was defined as the height difference between the corresponding node and its parent node:

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

Given the inherent signal-to-noise ratio of fMRI data, it is possible to observe that objects from different domains are clustered together (i.e. a node could include stimuli from multiple domains). Therefore, following the construction procedure of Xu and Tenenbaum (2007), we made two revisions to ensure that all nodes in the dendrogram could be utilized. First, the height of each node was scaled to lie between zero (for the lowest node)

and 0.5 (for the highest node). For the highest node, the prior probability was set 0.5, assuming that there was a virtual parent node with height 1, consistent with the largest height in Xu & Tenenbaum (2007).

*Likelihood probability computation.* In the word learning fMRI experiment (Experiment 2), each new word was associated with one or three exemplar objects. After the exemplars were given, the likelihood probability of each node is calculated as the reciprocal of the height of the corresponding node raised to a power of  $n$ , where  $n$  represents the number of exemplars ( $X$ ) of the current new word contained in that node (Equation 2):

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

If no exemplar is contained, the likelihood would be 0. A small constant value ( $\varepsilon$ ) was added to the height to avoid the likelihood probabilities being infinite at the lowest nodes.

*Posterior probability computation.* After the likelihood probabilities were calculated, the posterior probability of each node was then calculated using the Bayesian theorem (Equation 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 *generalization probability* of a probe object ( $y$ ) belonging to the current new word ( $W$ ) is calculated by summing the posterior probabilities of the nodes simultaneously containing both the probe object and any exemplar (Equation 4):

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

*Predicted neural representation computation.* To compute the neural representation of a learned concept (e.g. 'Leca'), we aggregated the measured object neural activity (in Experiment 1) of the same domain, weighted by the generalization probability of that object belonging to the concept. Fig 2 shows the flowchart of predicting the neural representations of a new word (e.g. 'Leca').

#### NMM

*Predicted neural representation computation.* To make predictions based on the NMM, the neural pattern of new word is expected to be the average of the t-value patterns from the corresponding exemplars. For example, for 'Leca', its predicted neural representation of the NMM was the averaged t-value pattern of the spoon, hammer and eyeglass, observed in Experiment 1.

*Generalization probability computation.* For a probe object, like the axe, its generalization probability belonging to 'Leca' was calculated as the Pearson correlation between the t-value pattern of the axe in Experiment 1 with the predicted one of 'Leca' of the NMM.

#### Prior-permuted NBM

Both construction and prediction of the prior-permuted NBM followed the same procedure as the NBM above, except that the objects associated with neural patterns were randomly shuffled when constructing the prior dendrogram.

#### BBM

Both construction and prediction of the BBM followed the same procedure as the NBM above, except that the prior dendrogram was constructed based on behavior-rating in the semantic distance-judgement task.

#### Model comparison

For each model, the predictive power of the neural representation of a given word was calculated as the Fisher-

transformed Pearson correlation coefficient between the predicted and observed neural pattern for each participant. For each participant and model, Fisher-z values were averaged across words referencing familiar objects ( $n = 12$ ) and novel shapes ( $n = 3$ ), respectively. We conducted paired samples t-tests to compare the predictive power of the NBM and NMM in learning with familiar objects and to compare the predictive power of the NBM between learning with familiar objects and with novel shapes. Partial correlation analyses were performed to test whether the NBM had additional power beyond that of the NMM and BBM. To this end, for each participant and word, the Fisher-transformed Pearson correlation coefficient was calculated between the predicted neural pattern of the NBM and the observed one was calculated, with the predicted neural pattern of the NMM or BBM was controlled for, respectively. The Fisher-z values were then averaged across word concept referencing familiar objects and entered into a one-sample t-test. A permutation test was performed to test whether the NBM outperformed the prior-permuted NBM, in which the prior in the NBM was permuted 100 times. The null hypothesis distribution was then constructed through bootstrap ( $n = 10,000$ ), in which one of the 100 permutations of each word was randomly selected with replacement for each participant, and the correlations between the predicted and observed neural patterns were computed and then Fisher-transformed. The group-level averaged Fisher-z values in each bootstrap formed the null hypothesis distribution. To provide a quantitative measure of the magnitude across effects and regions, we calculated the SES as  $\frac{(x-\mu)}{\sigma}$ , where  $x$  is the observed mean value,  $\mu$  is the mean of the null distribution, and  $\sigma$  is the standard deviation of the null distribution (Botta-Dukát, 2018). The right-tail P-value was estimated by approximating a standard normal distribution to the null distribution.

The predictive power of each model for generalization behavior was calculated using Pearson correlation analysis between the predicted and observed generalization probabilities across all probe objects of each new word. Steiger's Z-test was performed to compare the predictive power of the NBM and NMM. The partial correlation analysis was performed to test whether the NBM had additional power beyond the NMM and BBM. A permutation test was performed to test whether the NBM outperformed the prior-permuted NBM, which was repeated 1,000 times. The correlation coefficients across these permutations form the null hypothesis distribution, based on which the SES for the observed correlation coefficient was computed. The right-tailed P-value was estimated in the same manner as for predicting neural representations. Fisher-z test was performed to compare the predictive power of the NBM between learning with familiar objects and with novel shapes.

### ***Second-level analysis***

In the 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 pattern of the NMM and BBM were controlled. These maps were then smoothed using a 6 mm FWHM Gaussian kernel for subsequent second-level statistical analyses, conducting the above model comparison analysis using SPM12. Multiple comparison corrections were conducted using cluster-level FWE correction ( $P < .05$ ) as implemented in SPM12 (voxel-wise  $P < .001$ ). For the whole-brain level permutation test, we performed it in each of 96 cortical regions in Harvard-Oxford Atlas instead of at voxel level due to computation constraints, and multiple comparisons were corrected across the entire brain regions using the false-discovery rate (FDR) correction algorithm ( $q < 0.05$ ).

### **Supplementary Materials**

Supplemental Materials include the behavior data analyses and results (Section A) and the supplemental fMRI results (Section B).

### **Author contributions**

G.Z and Y.B. conceptualized and designed 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. revised the manuscript. Y.B. supervised the research.

### **Competing interests**

The authors declare that they have no conflict of interest. All authors approved the final version of the paper for submission.

### **Acknowledge**

This work was supported by the STI2030-Major Project (2021ZD0204104 to Y. Bi, 2022ZD0205104 to L. Zhu), National Natural Science Foundation of China (Grant No. 32400869 to G. Zhang; Grant No.32171052 to X. Wang; Grant No.31925020, 82021004 to Y. Bi; Grant No. 32071095 to L. Zhu), the China Postdoctoral Science Foundation (Grant No. 2023M740299, 2024T170062 to G. Zhang), the Fundamental Research Funds for the Central Universities (to Y. Bi), and Center for Life Sciences at Peking University (to L. Zhu). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the paper. We thank Jian Li for helpful comments on earlier drafts of the manuscript.

### **Data availability**

The data are available from Open Science Framework ([https://osf.io/wrt9s/?view\\_only=3f92f388670c4d44a00a0ca12dfccc4e](https://osf.io/wrt9s/?view_only=3f92f388670c4d44a00a0ca12dfccc4e)).

### **Code availability**

Custom code that supports the findings of this study is available from Open Science Framework ([https://osf.io/wrt9s/?view\\_only=3f92f388670c4d44a00a0ca12dfccc4e](https://osf.io/wrt9s/?view_only=3f92f388670c4d44a00a0ca12dfccc4e)).

## References

- Antony, J. W., Romero, A., Vierra, A. H., Luenser, R. S., Hawkins, R. D., & Bennion, K. A. (2022). Semantic relatedness retroactively boosts memory and promotes memory interdependence across episodes. *eLife*, 11, e72519. <https://doi.org/10.7554/eLife.72519>
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1), 95–113. <https://doi.org/10.1016/j.neuroimage.2007.07.007>
- Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, 26(3), 839–851. <https://doi.org/10.1016/j.neuroimage.2005.02.018>
- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., & others. (2025). Qwen2. 5-vl technical report. *arXiv Preprint arXiv:2502.13923*.
- Bi, Y., Wang, X., & Caramazza, A. (2016). Object Domain and Modality in the Ventral Visual Pathway. *Trends in Cognitive Sciences*, 20(4), 282–290. <https://doi.org/10.1016/j.tics.2016.02.002>
- Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3–4), 130–174. <https://doi.org/10.1080/02643294.2016.1147426>
- Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11), 527–536. <https://doi.org/10.1016/j.tics.2011.10.001>
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12), 2767–2796. <https://doi.org/10.1093/cercor/bhp055>
- Botta-Dukát, Z. (2018). Cautionary note on calculating standardized effect size (SES) in randomization test. *Community Ecology*, 19(1), 77–83. <https://doi.org/10.1556/168.2018.19.1.8>
- Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3), 389–414. <https://doi.org/10.1037/a0026450>
- Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *eLife*, 9. <https://doi.org/10.7554/eLife.59360>
- Bowman, C. R., & Zeithamova, D. (2018). Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 38(10), 2605–2614. <https://doi.org/10.1523/JNEUROSCI.2811-17.2018>
- Cameron, K. A., Yashar, S., Wilson, C. L., & Fried, I. (2001). Human Hippocampal Neurons Predict How Well Word Pairs Will Be Remembered. *Neuron*, 30(1), 289–298. [https://doi.org/10.1016/S0896-6273\(01\)00280-X](https://doi.org/10.1016/S0896-6273(01)00280-X)
- Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8), 354–361. [https://doi.org/10.1016/s1364-6613\(03\)00159-1](https://doi.org/10.1016/s1364-6613(03)00159-1)
- Carey, S. (1978). *The child as word learner*. <https://api.semanticscholar.org/CorpusID:17710641>
- Carey, S., & Bartlett, E. (1978). Acquiring a Single New Word. *Papers and Reports on Child Language Development*, 15, 17–29.
- Coutanche, M. N., & Thompson-Schill, S. L. (2015). Rapid consolidation of new knowledge in adulthood via fast mapping. *Trends in Cognitive Sciences*, 19(9), 486–488. <https://doi.org/10.1016/j.tics.2015.06.001>
- De Brigard, F., Umanath, S., & Irish, M. (2022). Rethinking the distinction between episodic and semantic memory: Insights from the past, present, and future. *Memory & Cognition*, 50(3), 459–463. <https://doi.org/10.3758/s13421-022-01299-x>
- Du, C., Fu, K., Wen, B., Sun, Y., Peng, J., Wei, W., Gao, Y., Wang, S., Zhang, C., Li, J., Qiu, S., Chang, L., & He, H. (2025). Human-like object concept representations emerge naturally in multimodal large language models. *Nature*

Machine Intelligence. <https://doi.org/10.1038/s42256-025-01049-z>

Duval, C., Desgranges, B., de La Sayette, V., Belliard, S., Eustache, F., & Piolino, P. (2012). What happens to personal identity when semantic knowledge degrades? A study of the self and autobiographical memory in semantic dementia. *Neuropsychologia*, 50(2), 254–265. <https://doi.org/10.1016/j.neuropsychologia.2011.11.019>

Fang, Y., Wang, X., Zhong, S., Song, L., Han, Z., Gong, G., & Bi, Y. (2018). Semantic representation in the white matter pathway. *PLoS Biology*, 16(4), e2003993. <https://doi.org/10.1371/journal.pbio.2003993>

Ferguson, B., & Waxman, S. (2017). Linking language and categorization in infancy. *Journal of Child Language*, 44(3), 527–552. <https://doi.org/10.1017/S0305000916000568>

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex (New York, N.Y. : 1991)*, 26(5), 2018–2034. <https://doi.org/10.1093/cercor/bhv020>

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, 81(2), 472–479. <https://doi.org/10.1111/j.1467-8624.2009.01408.x>

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(38), 15231–15235. <https://doi.org/10.1073/pnas.1221166110>

Fried, I., MacDonald, K. A., & Wilson, C. L. (1997). Single Neuron Activity in Human Hippocampus and Amygdala during Recognition of Faces and Objects. *Neuron*, 18(5), 753–765. [https://doi.org/10.1016/S0896-6273\(00\)80315-3](https://doi.org/10.1016/S0896-6273(00)80315-3)

Gabrieli, J. D., Cohen, N. J., & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain and Cognition*, 7(2), 157–177. [https://doi.org/10.1016/0278-2626\(88\)90027-9](https://doi.org/10.1016/0278-2626(88)90027-9)

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society*, 16(5), 748–753. Cambridge Core. <https://doi.org/10.1017/S1355617710000676>

Griffiths, T. L. (2024). Bayesian Models of Cognition. In M. C. Frank & A. Majid (Eds.), *Open Encyclopedia of Cognitive Science*. MIT Press.

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12, e82580. <https://doi.org/10.7554/eLife.82580>

Hirabayashi, T., Takeuchi, D., Tamura, K., & Miyashita, Y. (2013). Functional microcircuit recruited during retrieval of object association memory in monkey perirhinal cortex. *Neuron*, 77(1), 192–203. <https://doi.org/10.1016/j.neuron.2012.10.031>

Irish, M., Addis, D. R., Hodges, J. R., & Piguet, O. (2012). Considering the role of semantic memory in episodic future thinking: Evidence from semantic dementia. *Brain*, 135(7), 2178–2191. <https://doi.org/10.1093/brain/aws119>

Irish, M., & Piolino, P. (2016). Impaired capacity for prospection in the dementias – Theoretical and clinical implications. *British Journal of Clinical Psychology*, 55(1), 49–68. <https://doi.org/10.1111/bjc.12090>

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4), 169–188. Cambridge Core. <https://doi.org/10.1017/S0140525X10003134>

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9), 946–953. <https://doi.org/10.1038/78868>

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements.



940 *Frontiers in Psychology*, 3, 245. <https://doi.org/10.3389/fpsyg.2012.00245>

941 Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need?

942 Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, 20(7), 512–534.

943 <https://doi.org/10.1016/j.tics.2016.05.004>

944 Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic

945 program induction. *Science (New York, N.Y.)*, 350(6266), 1332–1338.

946 <https://doi.org/10.1126/science.aab3050>

947 Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like

948 people. *The Behavioral and Brain Sciences*, 40, e253. <https://doi.org/10.1017/S0140525X16001837>

949 Lee, S. W., O'Doherty, J. P., & Shimojo, S. (2015). Neural computations mediating one-shot learning in the human

950 brain. *PLoS Biology*, 13(4), e1002137. <https://doi.org/10.1371/journal.pbio.1002137>

951 LUPYAN, G. (2006). LABELS FACILITATE LEARNING OF NOVEL CATEGORIES. In *The Evolution of Language* (1–0, pp.

952 190–197). WORLD SCIENTIFIC. [https://doi.org/10.1142/9789812774262\\_0025](https://doi.org/10.1142/9789812774262_0025)

953 Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate

954 learning of novel categories. *Psychological Science*, 18(12), 1077–1083. [https://doi.org/10.1111/j.1467-](https://doi.org/10.1111/j.1467-9280.2007.02028.x)

955 [9280.2007.02028.x](https://doi.org/10.1111/j.1467-9280.2007.02028.x)

956 Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects

957 new conceptual knowledge. *Proceedings of the National Academy of Sciences of the United States of*

958 *America*, 113(46), 13203–13208. <https://doi.org/10.1073/pnas.1614048113>

959 Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58, 25–45.

960 <https://doi.org/10.1146/annurev.psych.57.102904.190143>

961 McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into

962 complementary learning systems theory. *Journal of Experimental Psychology. General*, 142(4), 1190–1210.

963 <https://doi.org/10.1037/a0033812>

964 McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010).

965 Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in*

966 *Cognitive Sciences*, 14(8), 348–356. <https://doi.org/10.1016/j.tics.2010.06.002>

967 Naya, Y. (2016). Declarative association in the perirhinal cortex. *Neuroscience Research*, 113, 12–18.

968 <https://doi.org/10.1016/j.neures.2016.07.001>

969 Naya, Y., Sakai, K., & Miyashita, Y. (1996). Activity of primate inferotemporal neurons related to a sought target in

970 pair-association task. *Proceedings of the National Academy of Sciences*, 93(7), 2664–2669.

971 <https://doi.org/10.1073/pnas.93.7.2664>

972 Naya, Y., Yoshida, M., Takeda, M., Fujimichi, R., & Miyashita, Y. (2003). Delay-period activities in two subdivisions of

973 monkey inferotemporal cortex during pair association memory task. *The European Journal of Neuroscience*,

974 18(10), 2915–2918. <https://doi.org/10.1111/j.1460-9568.2003.03020.x>

975 O'Kane, G., Kensinger, E. A., & Corkin, S. (2004). Evidence for semantic learning in profound amnesia: An

976 investigation with patient H.M. *Hippocampus*, 14(4), 417–425. <https://doi.org/10.1002/hipo.20005>

977 Open AI. (2024). Gpt-4o system card. *arXiv Preprint arXiv:2410.21276*.

978 Peelen, M. V., & Downing, P. E. (2005). Selectivity for the Human Body in the Fusiform Gyrus. *Journal of*

979 *Neurophysiology*, 93(1), 603–608. <https://doi.org/10.1152/jn.00513.2004>

980 Perszyk, D. R., & Waxman, S. R. (2019). Infants' advances in speech perception shape their earliest links between

981 language and cognition. *Scientific Reports*, 9(1), 3293. <https://doi.org/10.1038/s41598-019-39511-9>

982 Postle, B. R., & Corkin, S. (1998). Impaired word-stem completion priming but intact perceptual identification

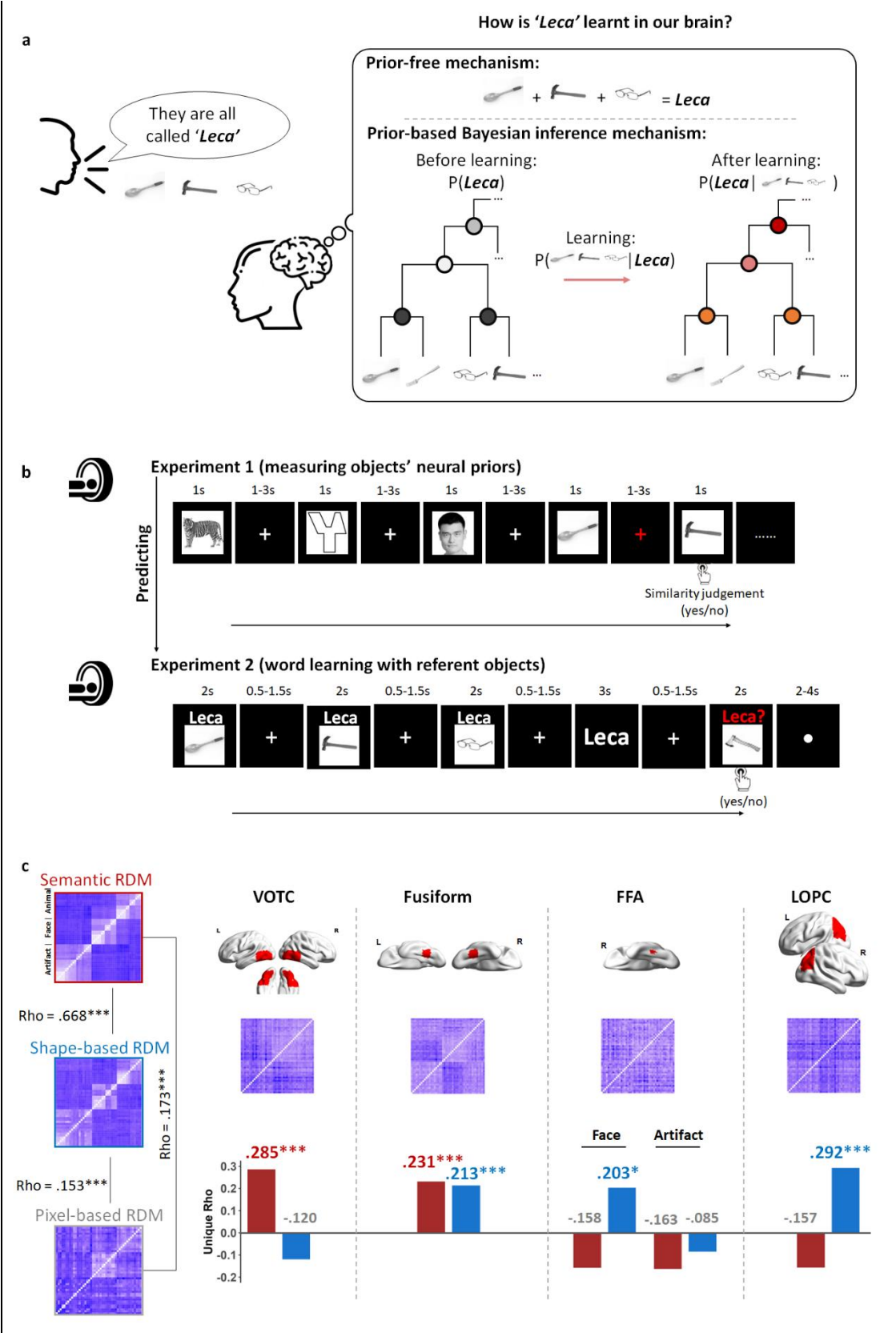
983 priming with novel words: Evidence from the amnesic patient H.M. *Neuropsychologia*, 36(5), 421–440.

[https://doi.org/10.1016/s0028-3932\(97\)00155-3](https://doi.org/10.1016/s0028-3932(97)00155-3)

- Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8), 587–597. <https://doi.org/10.1038/nrn3251>
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102–1107. <https://doi.org/10.1038/nature03687>
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.0) [Computer software]. <https://cran.r-project.org>
- Renoult, L., Irish, M., Moscovitch, M., & Rugg, M. D. (2019). From Knowing to Remembering: The Semantic–Episodic Distinction. *Trends in Cognitive Sciences*, 23(12), 1041–1057. <https://doi.org/10.1016/j.tics.2019.09.008>
- Renoult, L., & Rugg, M. D. (2020). An historical perspective on Endel Tulving’s episodic-semantic distinction. *Neuropsychologia*, 139, 107366. <https://doi.org/10.1016/j.neuropsychologia.2020.107366>
- Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, 76(4), 677–694. <https://doi.org/10.1016/j.neuron.2012.11.001>
- Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-024-01882-z>
- Sun, H., Zhao, L., Wu, Z., Gao, X., Hu, Y., Zuo, M., Zhang, W., Han, J., Liu, T., & Hu, X. (2024). Brain-like Functional Organization within Large Language Models. *arXiv Preprint arXiv:2410.19542*.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640. Cambridge Core. <https://doi.org/10.1017/S0140525X01000061>
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science (New York, N.Y.)*, 331(6022), 1279–1285. <https://doi.org/10.1126/science.1192788>
- Theves, S., Neville, D. A., Fernández, G., & Doeller, C. F. (2021). Learning and Representation of Hierarchical Concepts in Hippocampus and Prefrontal Cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 41(36), 7675–7686. <https://doi.org/10.1523/JNEUROSCI.0657-21.2021>
- Tulving, E. (1972). “Episodic and semantic memory,” in Organization of Memory. (*No Title*), 381.
- van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: From congruent to incongruent. *Neuropsychologia*, 51(12), 2352–2359. <https://doi.org/10.1016/j.neuropsychologia.2013.05.027>
- van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 30(47), 15888–15894. <https://doi.org/10.1523/JNEUROSCI.2674-10.2010>
- van Kesteren, M. T. R., Rijpkema, M., Ruiter, D. J., Morris, R. G. M., & Fernández, G. (2014). Building on prior knowledge: Schema-dependent encoding processes relate to academic performance. *Journal of Cognitive Neuroscience*, 26(10), 2250–2261. [https://doi.org/10.1162/jocn\\_a\\_00630](https://doi.org/10.1162/jocn_a_00630)
- Walsh, C. R., & Rissman, J. (2023). Behavioral representational similarity analysis reveals how episodic learning is influenced by and reshapes semantic memory. *Nature Communications*, 14(1), 7548. <https://doi.org/10.1038/s41467-023-42770-w>
- Wang, Y., Mao, X., Li, B., Lu, B., & Guo, C. (2016). Semantic memory influences episodic retrieval by increased familiarity. *NeuroReport*, 27(10). [https://journals.lww.com/neuroreport/fulltext/2016/07010/semantic\\_memory\\_influences\\_episodic\\_retri](https://journals.lww.com/neuroreport/fulltext/2016/07010/semantic_memory_influences_episodic_retri)

eval\_by.10.aspx

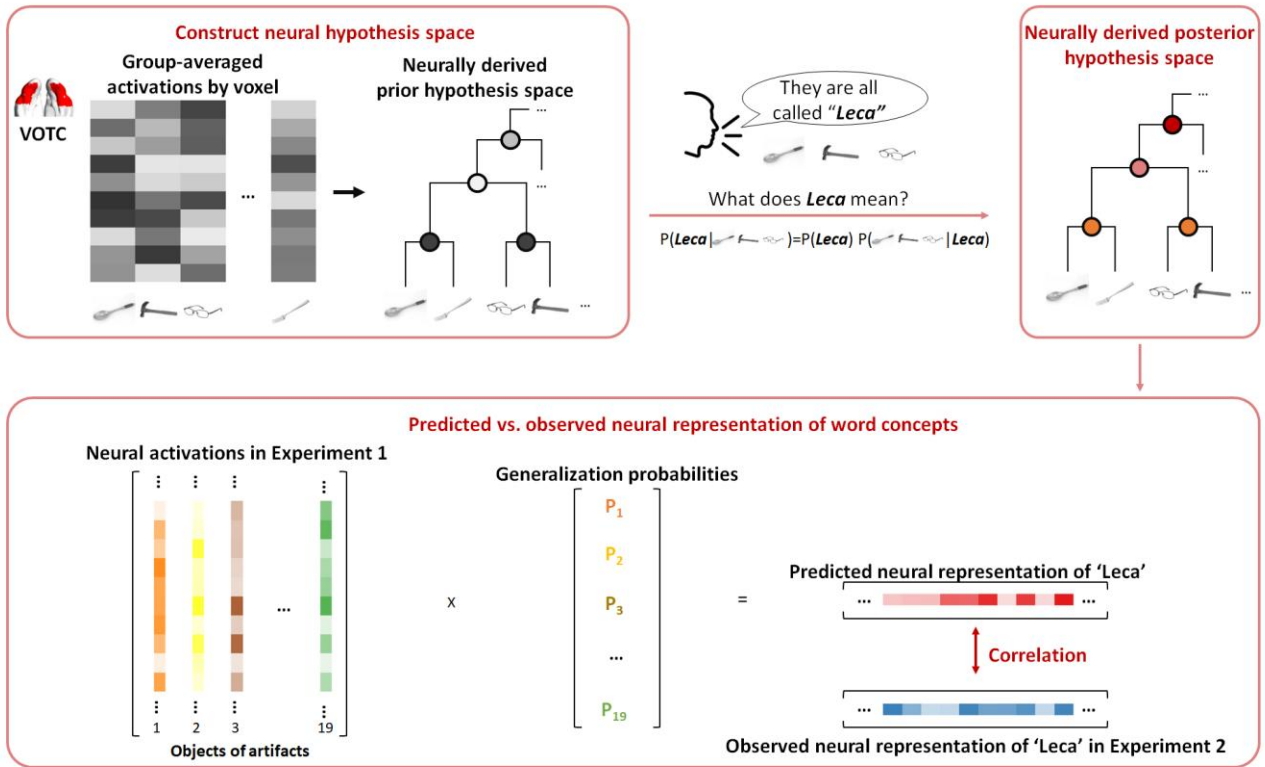
- Warren, D. E., & Duff, M. C. (2014). Not so fast: Hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus*, 24(8), 920–933. <https://doi.org/10.1002/hipo.22279>
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3), 257–302. <https://doi.org/10.1006/cogp.1995.1016>
- Wu, C. M., Meder, B., & Schulz, E. (2024). Unifying Principles of Generalization: Past, Present, and Future. *Annual Review of Psychology*. <https://doi.org/10.1146/annurev-psych-021524-110810>
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272. <https://doi.org/10.1037/0033-295X.114.2.245>
- Yan, C., & Zang, Y. (2010). DPARSF: a MATLAB toolbox for ‘pipeline’ data analysis of resting-state fMRI. *Frontiers in Systems Neuroscience*, 4. <https://www.frontiersin.org/articles/10.3389/fnsys.2010.00013>
- Yan, C.-G., Wang, X.-D., Zuo, X.-N., & Zang, Y.-F. (2016). DPABI: Data Processing & Analysis for (Resting-State) Brain Imaging. *Neuroinformatics*, 14(3), 339–351. <https://doi.org/10.1007/s12021-016-9299-4>
- Yanike, M., Wirth, S., Smith, A. C., Brown, E. N., & Suzuki, W. A. (2009). Comparison of Associative Learning-Related Signals in the Macaque Perirhinal Cortex and Hippocampus. *Cerebral Cortex*, 19(5), 1064–1078. <https://doi.org/10.1093/cercor/bhn156>



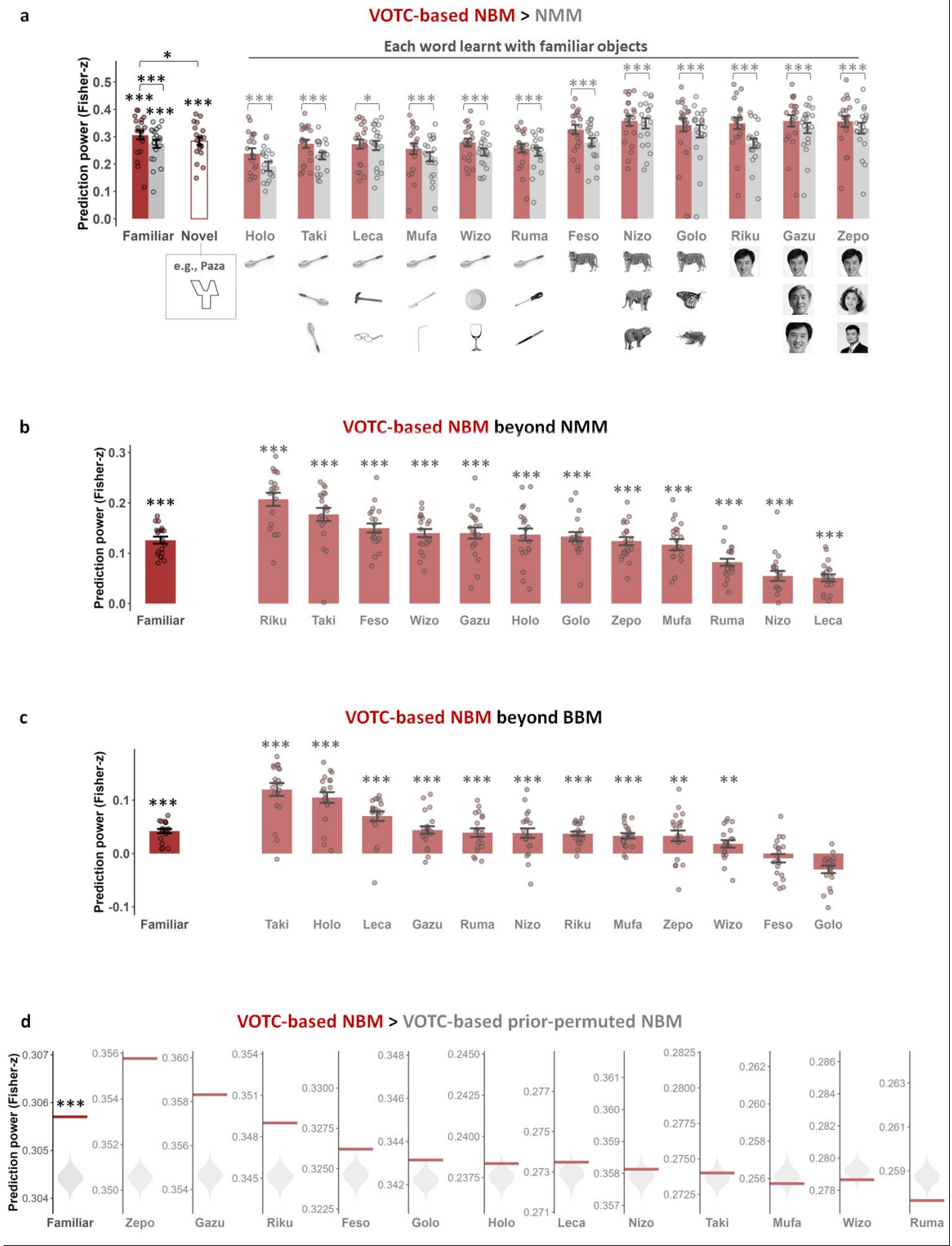
**a.** An example illustrates the core question in the present study: how a new word (e.g. ‘Leca’) is learned in our brain via reference to few exemplars (e.g. the spoon, hammer, and eyeglass). Two candidate mechanisms are proposed. The prior-free mechanism posits that the new word is formed based on only the exemplars, while the prior-based Bayesian inference mechanism argues that the structured priors construct the hypothesis space and that learning a new word engages updating this hypothesis space given the exemplars via Bayesian inference.

**b.** Participants (N=20) needed to complete two experiments in order undergoing fMRI scanning. In Experiment 1 (oddball one-back similarity judgement task), which is independent from learning, participants needed to view each object and judge whether the object after red fixation point was similar with last object (1s stimulus, 1–3s fixation). In Experiment 2 (word learning task), participants needed to learn a new word via reference to few exemplars (2s presentation, 0.5–1.5s fixation) and then judge whether the new word could be generalized to the probe object (2s judgment phase). The key of this two-phase design is that we used the neural representations of each object obtained from Experiment 1 to predict both the neural representations of new words and the generalization behavior obtained from Experiment2.

**c.** Neural priors contain more complex information beyond behavior-rating priors. The behavior-semantic (red outline), behavior-shape-based (blue outline), and pixel-controlled (gray outline) representational dissimilarity matrices (RDMs) were correlated with neural RDMs in different brain regions. Partial correlations (controlling for competing RDMs) reveal distinct associations over a distributed network: higher-order ventral visual cortex (ventral occipitotemporal cortex, VOTC) correlates overall with the semantic RDM; bilateral fusiform (within the VOTC) correlates with both the semantic and shape-based RDMs; fusiform face area (FFA, within the VOTC) correlates selectively with the face-shape-based RDM; lateral occipitoparietal cortex (LOPC) correlates with the shape-based RDM. Brain regions were defined using the Harvard-Oxford Atlas, except that the fusiform face area was defined by contrasting face pictures against animal and artifact pictures in Experiment 1 (120 top voxels around the peak voxel [40, -40, -22]; voxel-wise  $T_s > 4.870$ ). \*\*\* $P < .001$ ; \* $P < .05$ .



**Fig 2.** This flowchart illustrates how the Neural Bayesian Model (NBM) is constructed and how neural representation of 'Leca' is predicted via the NBM. The key procedure is to construct the NBM based on the group-averaged activations obtained in Experiment 1, then to predict the observed new word neural representation in Experiment 2. To construct the NBM in a given brain region, a hierarchical analysis is performed based on the similarities between the group-averaged activation pattern (t values) of each object stimuli. This analysis results in a dendrogram representing the neurally derived prior hypothesis space (Fig S5 shows the whole dendrogram based on the neural priors in the VOTC and on the behavioral priors). Each node in the dendrogram represents a possible hypothesis. Given the exemplars in learning, the prior hypothesis space is updated into the posterior one via Bayesian inference, according to which the probability of a probe object belonging to Leca, i.e. the generalization probability, is predicted. 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 corresponding generalization probabilities. The correlations between the predicted and observed neural representations of new words are calculated to evaluate the predictive power of the NBM, with higher correlations indicating greater predictive power.



**Fig 3.** VOTC-based NBM predicts the new word neural representations in the VOTC.

**a.** VOTC-based NBM outperforms the neural mean model (NMM, based on the prior-free mechanism) in learning words with familiar objects. In learning with novel shapes, although the VOTC-based NBM shows significant predictions (the white bar outlined in red), it is still worse than the predictions of the same model in learning with familiar objects, indicating that the NBM specifically applies to learning with rich priors. We also show the comparisons between the VOTC-based NBM and the NMM in each word learnt with familiar objects. The advantages of the VOTC-based NBM over the NMM exist in all words. The exemplars of each word are shown below the corresponding bar. In human face domain, to avoid copyright and related issues, AI-generated faces are used to replace real stimuli. Each point in each bar-plot represents the result from one participant (the same below).

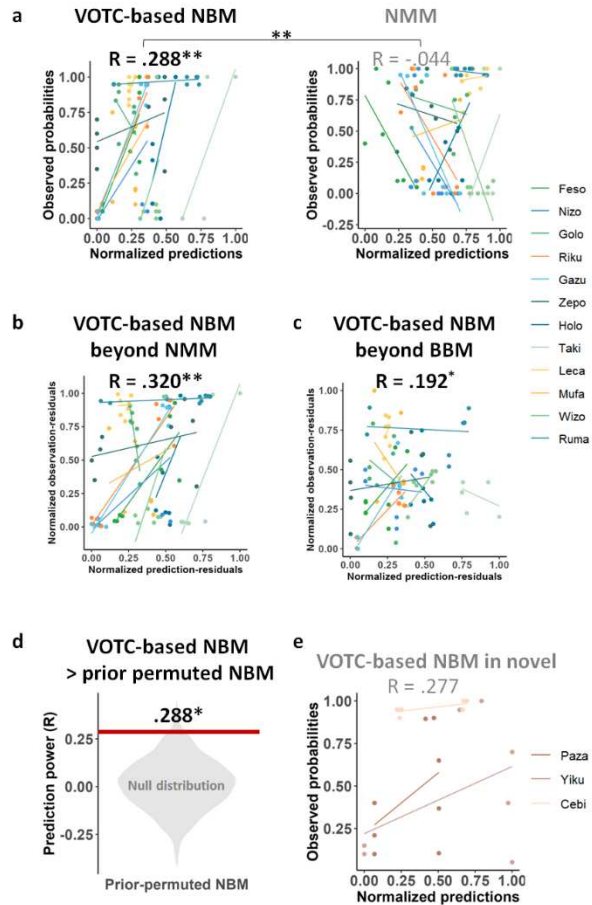
**b.** VOTC-based NBM has additional predictive power with the predictions of the NMM controlled for.

**c.** VOTC-based NBM has additional predictive power with the predictions of the behavioral Bayesian model (BBM) controlled for.

**d.** VOTC-based NBM outperforms the prior-permuted NBM. The violins represent the null distributions and the red lines represent the observed group-averaged predictive power of the NBM.

\*\*\*P < .001; \*\*P < .01, \*P < .05.

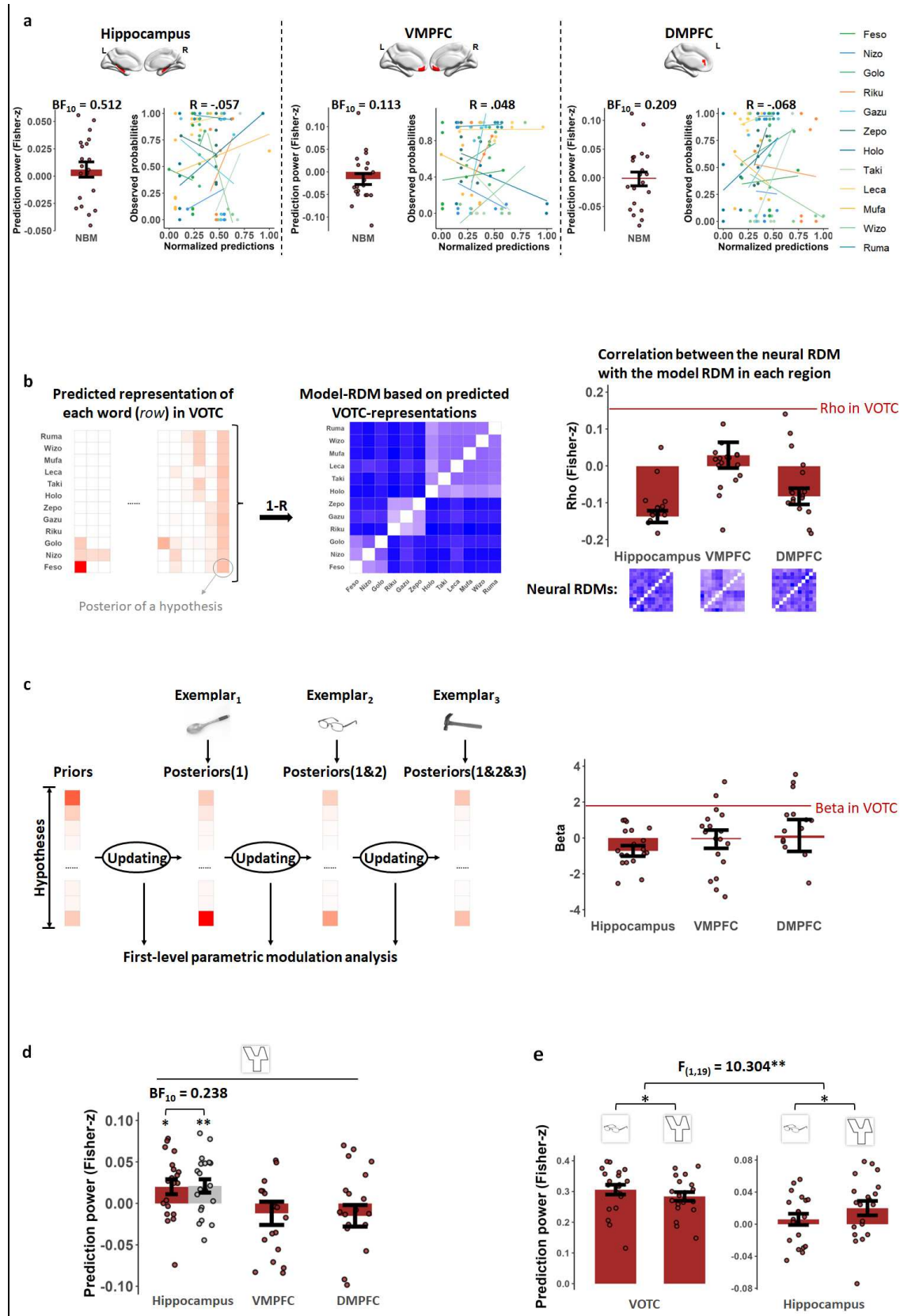




**Fig 4.** VOTC-based NBM predicts the generalization behavior in learning with familiar objects.

- a.** VOTC-based NBM outperforms the NMM in learning words with familiar objects. Each point indicates a unique trial (a combination between specific word and specific probe object) and trials in different words are indicated by different colors (the same below). The R-value above each scatterplot is the correlation across all trials, with nonsignificant results shown in gray. Predictions in each model are normalized across all trials to be 0-1.
- b.** VOTC-based NBM still has additional predictive power with the predictions of the NMM controlled for.
- c.** VOTC-based NBM still has additional predictive power with the predictions of the BBM controlled for.
- d.** VOTC-based NBM outperforms the prior-permuted NBM.
- e.** VOTC-based NBM failed to predict generalization behavior in learning with novel shapes.

\*\*P < .01, \*P < .05.



**Fig 5.** Dissociable roles of the VOTC and the hippocampus in word learning driven by neural priors.

**a.** The NBM fails to predict new word neural representations in the hippocampus, VMPFC, and DMPFC (bar plots) or generalization behavior (scatter plots) in learning with familiar objects, using neural priors from corresponding

brain regions. Each point in each bar-plot represents the result from one participant. Each point in the scatterplot indicates a unique trial (a combination between specific word and specific probe object). Trials in different words are indicated by different colors.

**b.** The NBM fails to predict new word neural representations in the hippocampus, VMPFC, and DMPFC in learning with familiar objects, using neural priors from the VOTC. The model-RDM based on the VOTC-representations fails to correlate significantly with the neural RDMs in the hippocampus, VMPFC and DMPFC. The red line in the bar-plot indicates the result in the VOTC. Each point in the bar-plot represents the result from one participant.

**c.** The NBM fails to predict tracking of word-concept-representation-updating in the hippocampus, VMPFC, and DMPFC in learning with familiar objects, using neural priors from the VOTC. We quantify the updating using the VOTC-based NBM and fit it to the activation strength of the hippocampus, VMPFC and DMPFC via parametric modulation analysis, which reveals no significant results. The red line in the bar-plot indicates the result in the VOTC. Each point in the bar-plot represents the result from one participant.

**d.** The hippocampus supports novel-shape concept learning. The NBM predicts hippocampal neural representations in learning with novel shapes and shows no difference with the NMM (the gray bar and points). Each point in the bar-plot represents the result from one participant.

**e.** Double dissociation: the NBM shows better predictive power in learning with familiar objects than with novel shapes in the VOTC, but better predictive power in learning with novel shapes than with familiar objects in the hippocampus. Each point in the bar-plot represents the result from one participant.

**\*\*P < .01, \*P < .05.**

a Large language models to be tested:



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 response)’]



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 response)’]

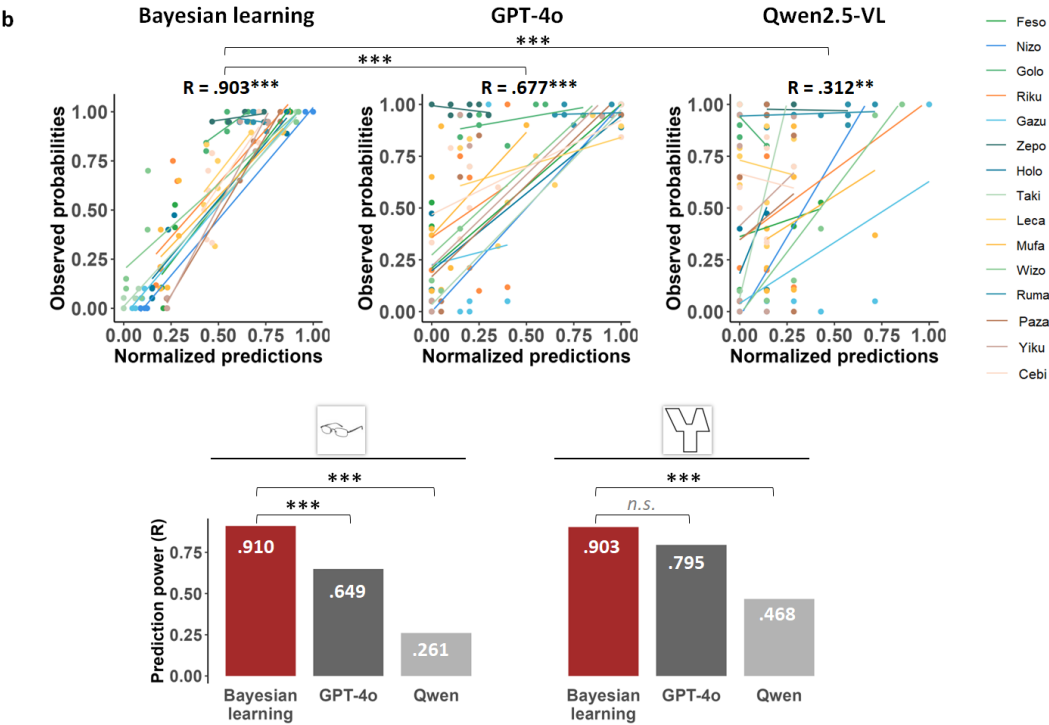


假如你有一位说另一种语言的朋友，这位朋友看到这一张图中的物体，把它称作'Leca' (无需回复)  
[‘Supposing that your foreign friend who speaks another language sees this object in this picture and calls it ‘Leca’ (no need to response)’]



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'.’]





**Fig 6.** Large language models (LLMs) are not as human-like as simple Bayesian learning model.  
a. We focus on two LLMs, including GPT-4o and Qwen2.5-VL, and ask them to perform the same learning task in humans. For a given trial, the exemplars are presented in three separate messages, mirroring the setting in human

experiments. All texts are prompted in Chinese to maintain consistency with the human learning task. Each trial is repeated 20 times in an independent chat session, setting temperature=1.0 and top-p=1.0 to reflect the original distribution predicted by the models. The generalization probability for a given trial was the proportion of positive responses over 20 iterations.

**b.** Overall, although both LLMs significantly predict generalization behavior, the Bayesian learning model outperforms them. These advantages persist in both learning with familiar objects and with novel shapes separately, except no significant difference between the Bayesian learning model with GPT-4o in learning with novel shapes. Here the predictions of the Bayesian learning model were obtained by fitting the observed generalization behavior against the predictions of both the VOTC-based NBM and the BBM via a general linear model. Predictions in each model are normalized across all trials to be 0-1. Each point in the scatterplot indicates a unique trial (a combination between specific word and specific probe object). Trials in different words are indicated by different colors.

\*\*\*P < .001; \*\*P < .01.

1163

1164

Tables

**Table 1.** Results of predicting human generalization behavior using LLMs and Bayesian learning model.

Learning condition	Models	R	95CI (lower boundary)	BF <sub>10</sub>	T	P	Compared with Bayesian learning model	
							z	P
Overall	GPT-4o	.677	.586	3.863E+14	9.998	< .001	5.833	< .001
	Qwen2.5-VL	.314	.171	1.414E+2	3.558	<.001	8.482	< .001
	Bayesian learning model*	.903	.870	4.919E+40	22.775	< .001	-	-
Familiar objects	GPT-4o	.649	.539	1.452E+10	8.269	< .001	6.711	< .001
	Qwen2.5-VL	.261	.096	10.542	2.621	0.005	9.115	< .001
	Bayesian learning model	.910	.876	3.009E+33	21.325	< .001	-	-
Novel Shapes	GPT-4o	.795	.620	4.762E+3	6.140	< .001	1.628	.104
	Qwen2.5-VL	.468	.148	7.531	2.485	.011	3.705	< .001
	Bayesian learning model	.903	.812	1.568E+6	9.883	< .001	-	-

1165

1166

1167

1168

*Note.* The degree of freedom was 118, 94 and 22 for learning overall, learning with familiar objects and learning with novel shapes. Abbreviations: 95CI = 95% confidence interval. The P value for ‘Compared with Bayesian learning model’ contrast is estimated according to Steiger’s Z-test (two-tailed). The remaining P values were right-tailed. \*The predictions of the Bayesian learning model were obtained by fitting the observed generalization behavior against the predictions of both the VOTC-based NBM and the BBM via general linear model.

## Supplementary Materials

### Section A: Behavioral analysis and results

#### *Participants showed high inter-subject correlation in two experiments*

##### Methods

For Experiment 1-2, because there was no absolute corrected answer, we calculated the inter-subject correlation (ISC) of participants' responses for each experiment. ISCs were calculated between each pair of participants using Equations 1, 2 and 3, where  $S_i$  and  $S_j$  were the responses of the  $i$ th and  $j$ th participants,  $I(S_i = S_j)$  was the number of the same responses between these two participants,  $n$  was the number of all responses for one participant,  $D$  was the absolute value of ISC, and  $Sign$  was the direction of ISC.

$$D = 1 - \frac{4 \times I(S_i = S_j) \times I(S_i \neq S_j)}{n^2} \quad (1)$$

$$Sign = \begin{cases} 1 & I(S_i = S_j) > \frac{n}{2} \\ 0 & I(S_i = S_j) = \frac{n}{2} \\ -1 & I(S_i = S_j) < \frac{n}{2} \end{cases} \quad (2)$$

$$ISC_{ij} = Sign \times D \quad (3)$$

In each experiment, the resulting ISCs were Fisher-transformed and then averaged. Null hypothesis significance testing was used to compare the averaged ISCs against chance. This was achieved via a permutation testing procedure. Specifically, for each task, the responses of each participant were randomly resampled from yes-or-no responses (10,000 times). On each occasion, we computed the resampling generated averaged 'ISCs'. These resampling ISCs formed empirical null distributions. We compared the actual ISCs against the empirical null distribution to compute the probability that the actual ISC was consistent with a chance ISC rate.

##### Results

ISCs (Fisher transformed) were 0.517 for Experiment 1 and 0.483 for Experiment 2. The bootstrap test showed that the 99.9% confidence intervals of chance level were  $[-0.003, 0.003]$  for each experiment. Therefore, both ISCs were higher than chance levels ( $P_s < .001$ ).

























#### *Replicating both previous behavior result patterns and the advantages of the BBM over behavioral mean model*

Participants' generalization behavior result patterns of the pilot experiment and Experiment 2 were shown in Fig S2. The results were highly consistent between two experiments ( $R = .955$ ,  $P < .001$ ). Both experiments replicated the result pattern of Xu & Tenenbaum (2007): when learning with a single exemplar, participants showed graded generalization to new objects that had high-, medium-, and low-level similarity with the learned exemplar; when learning with three exemplars, participants showed generalization sharpened into a much more all-or-none pattern depending on how similar the three exemplars were (Fig S2).

Given that only participants in Experiment 2 performed the semantic distance-judgement task, here we predicted the behavior results in Experiment 2 using the BBM and the behavioral mean model (BMM) based on behavioral ratings, to replicate the advantages of the BBM. To this end, for the Bayesian Model, we built a prior knowledge dendrogram based on the distance

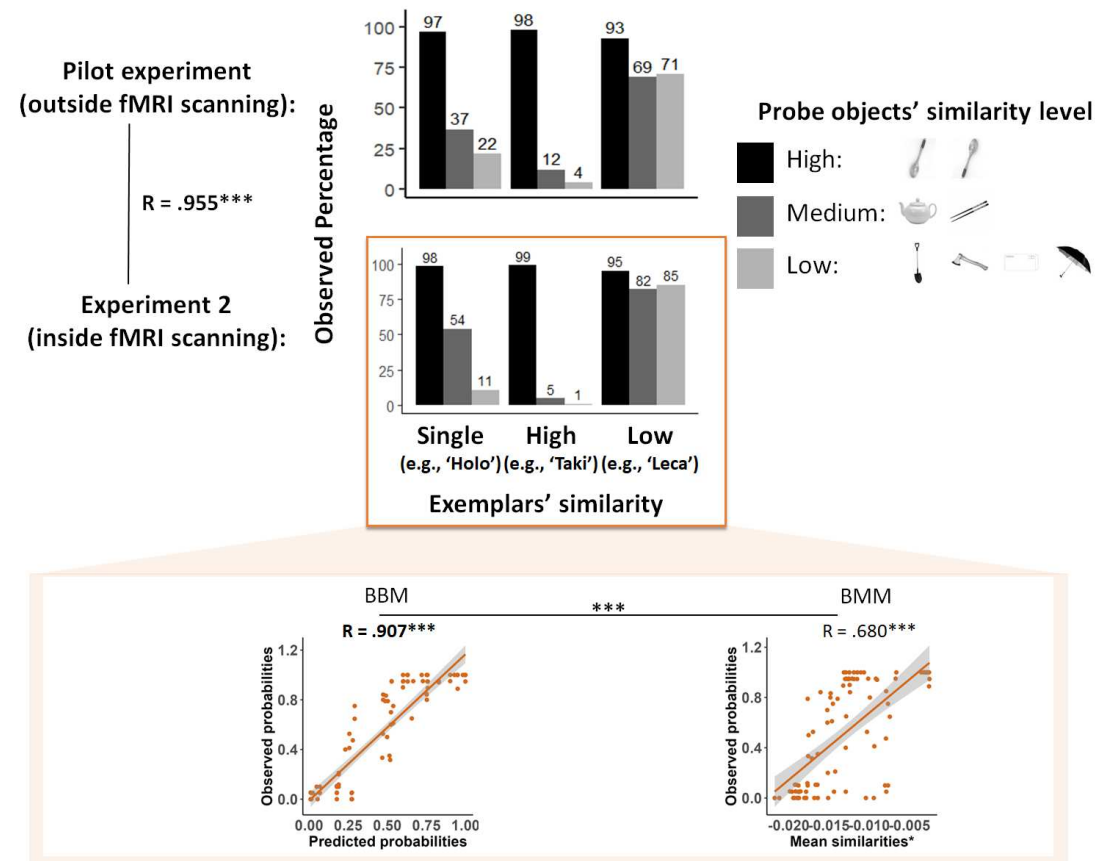
matrix resulting from the behavioral rating task. The probabilities of probes belonging to a given learned word were calculated following the same calculation principle as in ROI analysis. For the BMM, the probabilities of probe objects belonging to one given learned word were defined as the mean similarity (i.e, the opposite of the mean distance) between the probe objects and corresponding exemplars, with higher values indicating higher probabilities of the probe objects belonging to the corresponding words. The predictive power of each model was calculated as Pearson correlations between the predictive probabilities and observed probabilities (choosing percentages). The results replicated the advantages of the BBM over BMM (Fig S2).



Words	Exemplars	Probe objects
Feso		 
Nizo	  	 
Golo	  	   
Riku		 
Gazu	  	 
Zepo	  	   
Holo		
Taki	  	 
Leca	  	 
Mufa	  	   
Wizo	  	
Ruma	  	
Paza		 
Yiku	  	 
Cebi	  	   

**Fig S1.** All stimuli (including both exemplars and probe objects for all new words) used in Experiment 2. There are 58 unique objects, which correspond to the objects used in Experiment

1. In each domain, the eight probe objects are divided into three different groups according to their similarity to the exemplar outlined in gray. The two probe objects on the top row are of high similarity, the two ones on the middle row are of medium similarity, and the four ones on the bottom row are of low similarity. Similarity are defined according to the behavior rating in the semantic distance-judgement task.



**Fig S2.** Replicating the classical generalization behavior result pattern and the advantage of behavioral Bayesian model (BBM).

- The generalization behavior result patterns are consistent across pilot experiment and Experiment 2.
- Predictive power of the BBM outperforms the behavioral mean model (BMM) in predicting generalization behavior of humans. \*The mean similarities in the BMM are defined as the opposite of the mean distance between the probe objects and the exemplars of the given word.

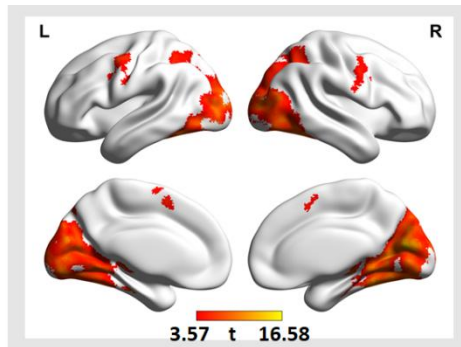
\*\*\*P < .001

**Section B: Supplemental fMRI results in the main article**

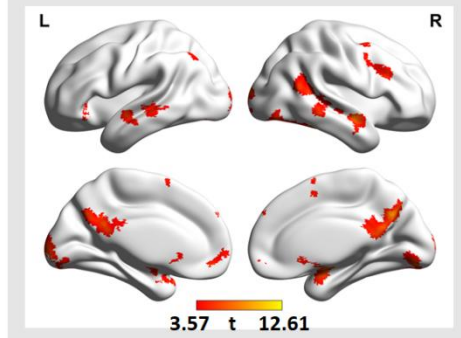
***Univariate activation of all sampled objects and of each domain in Experiment 1***

The sample objects activated regions in the bilateral temporooccipital regions extending to intraparietal sulcus, and bilateral precentral gyrus. Contrasting among the three domain of familiar objects revealed that face pictures showed stronger activities in the right posterior fusiform cortex, posterior cingulate gyrus, right lateral anterior temporal lobe, right posterior middle temporal gyrus, bilateral medial anterior temporal lobe, right middle temporal gyrus, left frontal orbital cortex, right middle frontal gyrus and ventral medial frontal cortex; animal pictures in regions extending from the bilateral posterior fusiform cortex to lateral occipital cortex; and artifact pictures in the bilateral lateral superior occipital cortex, cuneal cortex, and bilateral temporooccipital fusiform cortex. Compared to familiar objects, novel shape pictures strongly activated regions in the bilateral intraparietal cortex.

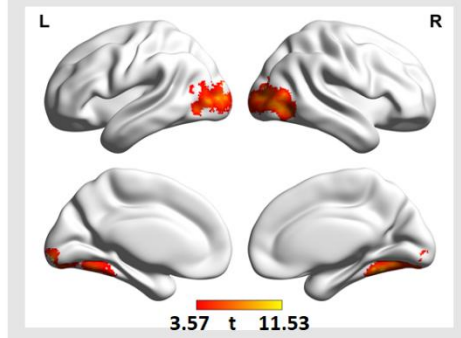
(Face + Animal + Artifact + Novel shapes) > 0



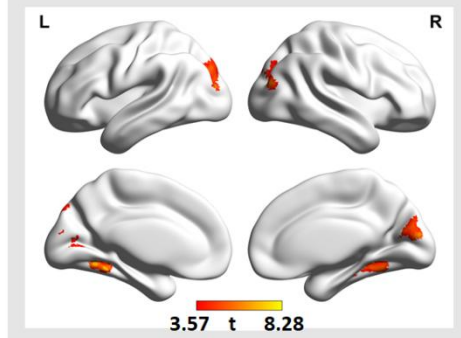
Face > (Animal + Artifact)



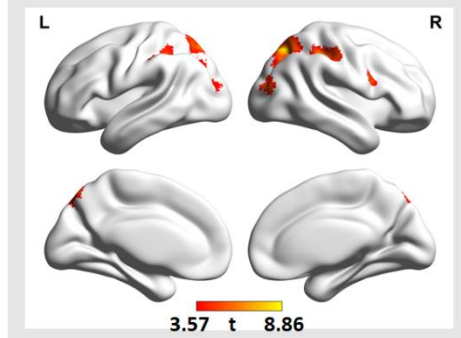
Animal > (Face + Artifact)



Artifact > (Animal + Face)



Novel shape > (Animal + Face + Artifact)



**Fig S3.** Whole-brain results of univariate activation of all sampled objects and of each domain.

1251

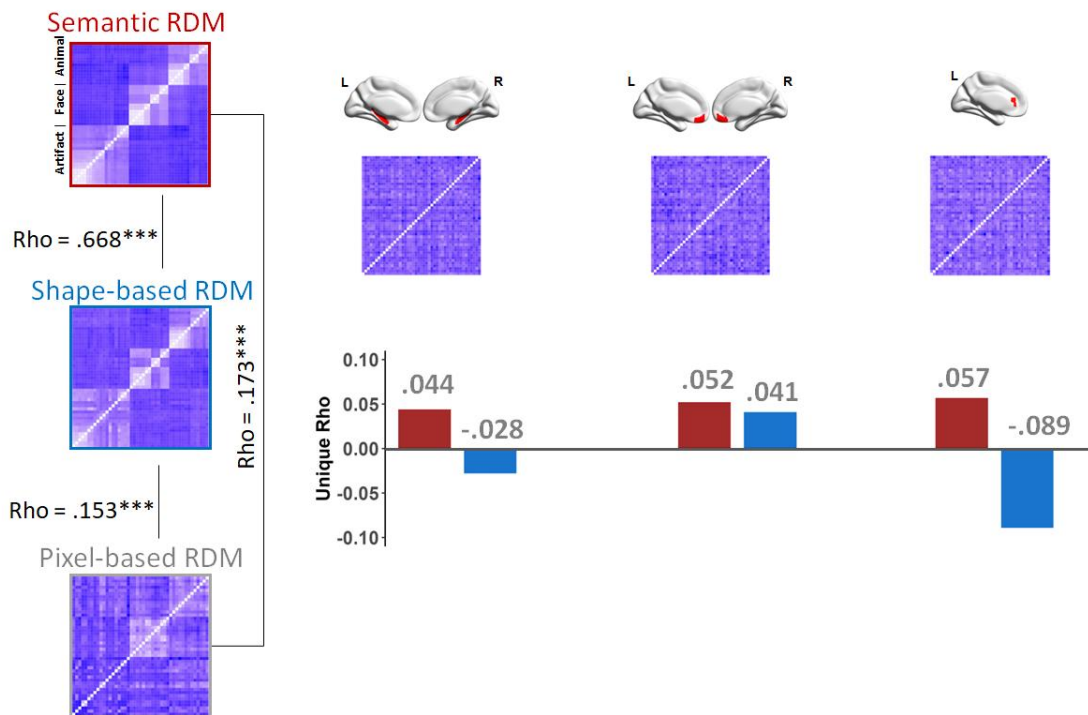
**Table S1**

1252

*Whole-brain results of univariate activation of all sampled objects and of each domain in Experiment 1 (voxel-wise  $P < .001$ , cluster-wise FWE  $P < .05$ ).*

Model contrasts	Anatomical region of the peak voxel	Number of voxels	MNI coordinates of the peak voxel			Peak T value
			x	y	z	
All sampled objects > 0						
	Lateral Occipital Cortex	25651	18	-88	0	16.583
	Left Juxtapositional Lobule Cortex (formerly Supplementary Motor Cortex)	314	-2	6	56	7.638
	Right Precentral Gyrus	744	52	-4	54	7.357
	Left Precentral Gyrus	701	-46	0	50	7.332
Face > (Animal +Artifact)						
	Right Superior Temporal Gyrus, anterior division	309	50	-6	-14	12.602
	Right Temporal Occipital Fusiform Cortex	2512	40	-40	-22	10.266
	Right Precuneous Cortex	1553	6	-62	34	9.522
	Right Temporal Pole	420	14	-4	-16	9.196
	Left Temporal Pole	159	-38	14	-26	7.536
	Left Frontal Medial Cortex	178	-2	50	-12	6.944
	Right Angular Gyrus	844	60	-38	-8	6.765
	Left Frontal Orbital Cortex	230	-40	22	-8	6.615
	Left Middle Temporal Gyrus, posterior division	263	-54	-30	-10	6.406
	Left Superior Temporal Gyrus, anterior division	251	-54	-8	-8	6.223
	Left Lateral Occipital Cortex, inferior division	101	-40	-76	-16	6.037
	Right Middle Frontal Gyrus	123	38	2	56	5.961
	Right Middle Frontal Gyrus	385	42	22	26	5.857

	Right Supplementary Motor Cortex	247	2	8	64	5.812
	Left Lateral Occipital Cortex, superior division	235	-38	-66	42	5.687
	Right Superior Frontal Gyrus	96	4	52	36	4.967
	Right Lateral Occipital Cortex, superior division	172	38	-58	40	4.939
Animal > (Face + Artifact)						
	Right Temporal Occipital Fusiform Cortex	3901	30	-54	-16	11.528
	Lateral Occipital Cortex	2886	-24	-84	-4	10.115
Artifact > (Animal + Face)						
	Left Temporal Occipital Fusiform Cortex	266	-26	-46	-14	8.275
	Left Lateral Occipital Cortex, superior division	899	-22	-78	26	7.716
	Right Cuneal Cortex	941	4	-86	16	6.307
	Right Temporal Occipital Fusiform Cortex	333	34	-42	-12	6.033
Novel shape > (Face + Animal +Artifact)						
	Right Lateral Occipital Cortex, superior division	1180	22	-70	50	8.855
	Right Precentral Gyrus	192	52	8	26	7.742
	Right Supramarginal Gyrus, posterior division	816	40	-38	42	7.178
	Left Lateral Occipital Cortex, superior division	846	-16	-74	54	6.649
	Right Lateral Occipital Cortex, superior division	376	30	-86	16	6.347
	Left Supramarginal Gyrus, anterior division	439	-48	-32	36	5.972
	Lateral Occipital Cortex	267	-22	-84	14	5.415



**Fig S4.** Neural RDMs of the hippocampus, VMPFC and DMPFC are not significantly correlated with either semantic or shape-based RDM.

\*\*\* $P < .001$ .

## **Whole-brain searchlight analysis results of examine effects of the NBM**

The whole-brain searchlight analyses consistently revealed the advantages of the NBM in predicting new word neural representations in the regions representing object knowledge, which were consistent with the results in the VOTC. Below are the detailed results for each analysis:

*Can the NBM predict new word neural representations?* In the whole-brain searchlight analysis, the NBM could predict neural representations in regions encompassing the bilateral fusiform gyrus, lateral occipitotemporal cortex (LOT), inferior parietal sulcus, supplementary motor area (SMA), precentral gyrus to middle frontal gyrus, and inferior frontal gyrus, which were consistently found to represent object knowledge (Fig S6 and Table S7).

*Does incorporating the neural priors matter?* Direct comparison revealed that the NBM significantly outperformed the NMM in predicting the neural representations of learned word in the aforementioned brain regions (Fig S6 and Table S7). No region showed advantages for the NMM. After controlling the NMM, the NBM still had unique predictions in the same regions (Fig S6 and Table S7). These supported the advantages of incorporating structured neural priors in the NBM.

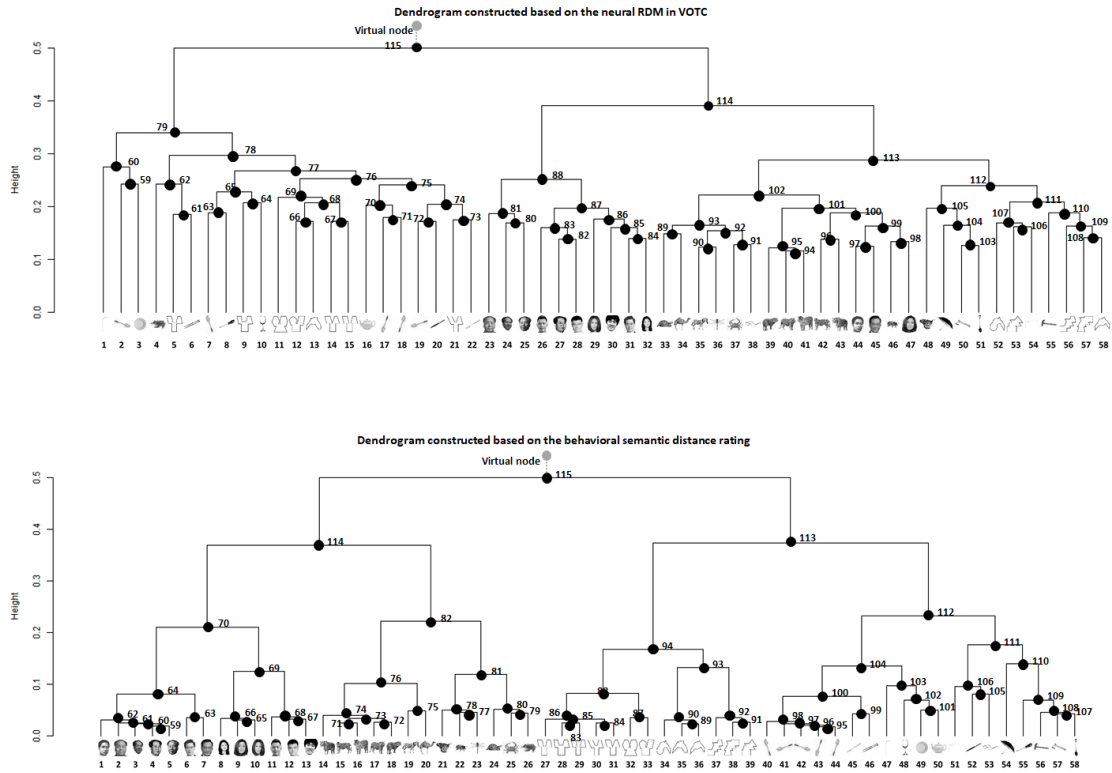
*Does incorporating the specific neural-priors matter?* Due to computation constraints, we performed this permutation test analysis on the brain region level instead of the voxel level, using the 96 cortical regions in Harvard-Oxford Atlas. The results were corrected for multiple comparisons across the 96 regions using the false-discovery rate (FDR) correction algorithm ( $q < 0.05$ ). As shown in Fig S6 (also see Table S8), the NBM outperformed prior-permuted control model in the bilateral temporooccipital fusiform cortex (subregions in the VOTC), left lateral occipitoparietal cortex, left occipital pole, right LOTC and left precuneus cortex. These regions were largely consistent with regions compared with 0 above (NBM > 0).

*Do neural priors have additional predictive power beyond behavioral priors?* After controlling the BBM, the NBM showed unique predictive power in the bilateral temporooccipital fusiform cortex (subregions in the VOTC), right occipital pole, SMA, left superior parietal lobe, left supramarginal gyrus extending to left postcentral gyrus (Fig S6 and Table S7). This supported that the structured neural priors had additional power beyond single behavioral semantics priors.

*Does the NBM specifically apply to word concept learning with rich priors?* Direct comparison between familiar objects and novel shapes revealed different preferences (Fig S6 and Table S7). In regions like the temporooccipital fusiform cortex, which is subregions in the VOTC and sensitive to multiple dimensions of information (Fig 1c), the NBM showed stronger predictive power in learning with familiar objects, while in regions like lateral occipitoparietal cortex, which is sensitive exclusively to the shape information, the NBM showed stronger efficacies in learning with novel shapes.

For generalization behavior, the whole-brain searchlight analysis failed to reveal any significant results for the NBM.





**Fig S5.** The whole dendrograms constructed based on the neural priors in the VOTC (upper) and on the behavioral priors (lower). In each dendrogram, different node (/hypothesis) are labelled by different numbers. It is noted that the two nodes have the same number in two dendrograms does not mean that they are the same hypothesis, given that the two dendrograms are not comparable.

1304 **Table S2.** ROI results of predicting both new word neural representations and generalization behavior in the VOTC, hippocampus, VMPFC and DMPFC using neural  
1305 priors from corresponding brain regions.

Dependent variables	ROI	Model contrasts	R*	95CI (lower boundary)	Cohen's d	BF <sub>10</sub>	T/SES*	P
New word neural representation	VOTC	NBM > 0	0.306 (0.016)	0.277	4.143	9.630E+10	18.529	< .001
		NMM > 0	0.274 (0.015)	0.248	4.091	7.753E+10	18.294	< .001
		NBM > NMM	0.031 (0.002)	0.027	3.724	7.934E+09	16.653	< .001
		NBM (with NMM controlled for) > 0	0.126 (0.007)	0.115	4.327	1.011E+11	19.353	< .001
		NBM > Prior-permuted NBM	-	-	-	-	7.569	< .001
		NBM (with BBM controlled for) > 0	0.042 (0.004)	0.034	2.246	4.851E+06	10.043	< .001
		NBM: Familiar > Novel	0.021 (0.008)	0.005	0.627	4.606	2.802	.011
Generalization behavior	Hippocampus	NBM > 0	0.006 (0.007)	-0.006	0.194	0.512	0.865	0.199
	VMPFC	NBM > 0	-0.016 (0.012)	-0.036	-0.287	0.113	-1.285	0.893
	DMPFC	NBM > 0	-0.002 (0.012)	-0.022	-0.033	0.209	-0.149	0.558
	VOTC	NBM > 0	.288	0.125	-	21.466	2.915	.002
		NMM > 0	-.044	-.242	-	0.173	-0.429	.666
		NBM > NMM	-	-	-	-	2.927	.003
		NBM (with NMM controlled for) > 0	.320	.159	-	54.923	3.270	<.001
		NBM > Prior-permuted NBM	-	-	-	-	2.048	.020
		NBM: Familiar > Novel	-	-	-	-	0.049	.961
		NBM (with BBM controlled for) > 0	.192	.024	-	2.372	1.894	.031
	Hippocampus	NBM > 0	-.057	-0.224	-	0.161	-0.552	.709
	VMPFC	NBM > 0	.048	-0.122	-	0.350	0.466	.321
	DMPFC	NBM > 0	-.068	-0.234	-	0.151	-0.657	.744

1306 *Note.* All contrasts except 'NBM: Familiar > Novel' were in learning with familiar objects. The degree of freedom was 19 for analyses of the new word neural  
1307 representation and 94 for analyses of generalization behavior. As a measure of the effect size, Cohen's d was calculated for each effect as the means divided by the

1308 pooled standard deviations. The P values for 'NBM > NMM' and 'NBM: Familiar > Novel' contrasts were two-tailed, and the rest P values were right-tailed. In predicting  
1309 generalization behavior, the P value for 'NBM > NMM' contrast is estimated according to Steiger's Z-test, and the P value for 'NBM: Familiar > Novel' is estimated  
1310 according to Fisher's Z-test. \*R values for new word neural representations were Fisher-transformed in the form of mean (standard error). \*The standard effect size  
1311 (SES) was calculated for each effect as the difference between the observed value and mean value of the null distribution, divided by the standard deviation of the  
1312 null distribution. Abbreviations: 95CI = 95% confidence interval; NBM = neural Bayesian model; NMM = neural mean model; BBM = behavioral Bayesian model.  
1313

1314 **Table S3**  
 1315 *ROI results of predicting new word neural representations in three domains of familiar objects using the NBM constructed based on neural priors from the VOTC.*

Domain	R (Fisher-Z)	95CI (lower boundary)	Cohen's d	BF <sub>10</sub>	T	P
Animal	0.342 (0.018)	0.311	4.157	1.020E+11	18.592	< .001
Face	0.354 (0.020)	0.320	4.004	5.388E+10	17.906	< .001
Artifact	0.263 (0.015)	0.236	3.818	2.417E+10	17.075	< .001

1316 *Note.* The degree of freedom was 19 for all analyses. R values were Fisher-transformed in the form of mean (standard error). As a measure of the effect size,  
 1317 Cohen's d was calculated for each effect as the means divided by the pooled standard deviations. Abbreviations: 95CI = 95% confidence interval. All P values were  
 1318 right-tailed.  
 1319

1320 **Table S4.** ROI RSA results of predicting neural patterns of new words in the VOTC, hippocampus, VMPFC and DMPFC by the representations in the VOTC-based NBM.

ROI	Rho (Fisher-Z)	95CI (lower boundary)	Cohen's d	BF <sub>10</sub>	T	P
VOTC	0.156 (0.039)	0.088	0.894	92.148	4.000	< .001
Hippocampus	-0.137 (0.016)	-0.166	-1.882	0.049	-8.416	> .999
VMPFC	0.030 (0.035)	-0.031	0.191	0.505	0.854	.202
DMPFC	-0.082 (0.022)	-0.12	-0.826	0.064	-3.692	.999

1321 *Note.* Learning was based on familiar objects. The degree of freedom was 19 for all analyses. Spearman-rho values were Fisher-transformed in the form of mean  
1322 (standard error). As a measure of the effect size, Cohen's d was calculated for each effect as the means divided by the pooled standard deviations. Abbreviations:  
1323 95CI = 95% confidence interval; NBM = neural Bayesian model. All P values were right-tailed.

1324  
1325 **Table S5.** ROI results of testing whether activities of the VOTC, hippocampus, VMPFC and DMPFC can track neural representation updating in the VOTC-based NBM.

ROI	Beta	95CI (lower boundary)	Cohen's d	BF <sub>10</sub>	T	P
VOTC	1.794 (0.903)	0.233	0.444	2.275	1.987	0.031
Hippocampus	-0.721 (0.295)	-1.231	-0.546	0.079	-2.444	0.988
VMPFC	-0.069 (0.509)	-0.949	-0.03	0.211	-0.135	0.553
DMPFC	0.139 (0.887)	-1.394	0.035	0.262	0.157	0.438

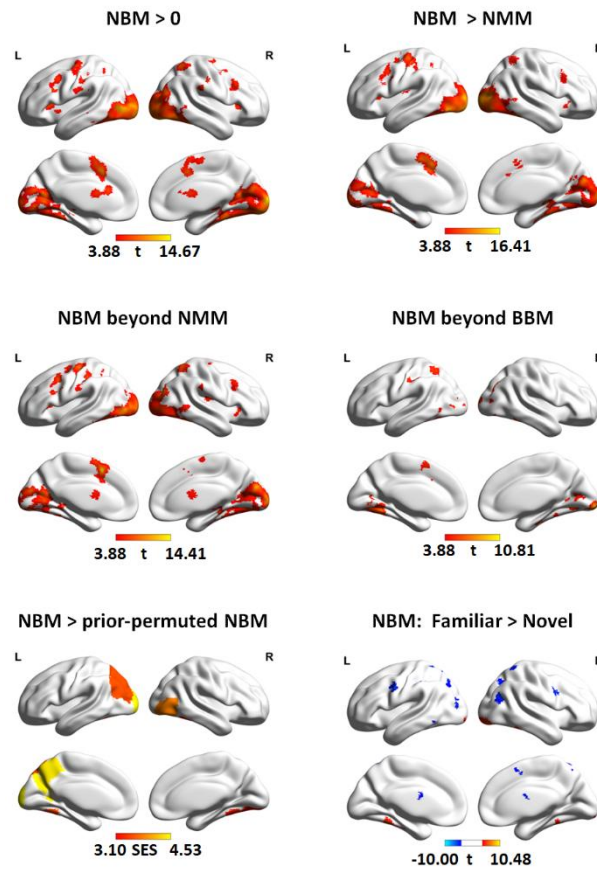
1326 *Note.* Learning is based on familiar objects. The degree of freedom was 19 for all analyses. Beta values were in the form of mean (standard error). As a measure of  
1327 the effect size, Cohen's d was calculated for each effect as the means divided by the pooled standard deviations. Abbreviations: 95CI = 95% confidence interval;  
1328 NBM = neural Bayesian model. All P values were right-tailed.

1329

1330 **Table S6.** ROI results of new word neural representations in the hippocampus, VMPFC and DMPFC in learning with novel shapes, using neural priors from  
1331 corresponding brain regions.

ROI	Model contrasts	R	95CI (lower boundary)	Cohen's d	BF <sub>10</sub>	T	P
Hippocampus	NBM > 0	0.020 (0.009)	0.006	0.534	4.399	2.389	.014
	NMM > 0	0.021 (0.008)	0.007	0.593	6.97	2.652	.008
	NBM > NMM	-0.001 (0.003)	-0.007	-0.053	0.238	-0.236	.816
	NBM (with NMM controlled for) > 0	0.009 (0.006)	-0.001	0.354	1.247	1.581	.065
	NMM (with NBM controlled for) > 0	0.002 (0.006)	-0.008	0.066	0.294	0.293	.386
	NBM: Novel > Familiar	0.015 (0.006)	0.001	0.504	1.794	2.254	.036
	NMM: Novel > Familiar	0.019 (0.007)	0.004	0.586	3.325	2.620	.017
VMPFC	NBM > 0	-0.012 (0.014)	-0.036	-0.18	0.141	-0.806	.785
DMPFC	NBM > 0	-0.015 (0.013)	-0.037	-0.255	0.12	-1.141	.866

1332 *Note.* Degree of freedom was 19 for all analyses. R values were Fisher-transformed with the form of mean (standard error). As a measure of effect size, Cohen's d  
1333 was calculated for each effect as the means divided by the pooled standard deviations. Abbreviations: 95CI = 95% confidence interval; NBM = neural Bayesian  
1334 model; NMM = neural mean model. The P values for 'NBM > NMM', 'NBM: Novel > Familiar' and 'NMM: Novel > Familiar' contrasts were two-tailed, and the rest P  
1335 values were right-tailed.



**Fig S6.** Whole-brain searchlight results of predicting new word neural representations. All analyses followed the same methods as that in ROI analysis. Abbreviations: NBM = neural Bayesian model; NMM = neural mean model; BBM = behavioral Bayesian model.

1340

**Table S7**

1341

*Whole-brain searchlight results of predicting new word neural representations (voxel-wise  $P < .001$ , cluster-wise FWE  $P < .05$ ).*

Model contrasts	Anatomical region of the peak voxel	Number of voxels	MNI coordinates of the peak voxel			Peak T value
			x	y	z	
NBM > 0						
	Right Occipital Pole	30514	10	-102	6	14.664
	Left Precentral Gyrus	5836	-32	8	38	12.286
	Right Postcentral Gyrus	107	60	-18	42	8.495
	Right Superior Parietal Lobule	376	30	-56	60	8.442
	Right Middle Frontal Gyrus	370	42	4	60	8.136
	Right Insular Cortex	453	26	34	6	7.663
	Right Supramarginal Gyrus	152	60	-30	28	7.590
	Right Lateral Superior Occipital Cortex	146	38	-60	44	6.761
	Right Middle Frontal Gyrus	295	46	20	34	6.316
	Right Precentral Gyrus	116	26	-12	66	6.178
	Cingulate Gyrus	103	22	16	28	5.821
	Left Postcentral Gyrus	102	-64	-22	28	5.108
NBM > NMM						
	Right Occipital Pole	20115	34	-96	6	16.393
	Left Postcentral Gyrus	1347	-52	-22	58	10.748
	Left Paracingulate Gyrus	1772	-6	16	44	10.694
	Left Superior Parietal Lobule	387	-20	-50	42	10.609
	Right Paracingulate Gyrus	326	12	14	50	8.977
	Right Insular Cortex	142	30	26	4	8.565



	Left Insular Cortex	252	-30	22	6	8.438
	Right Superior Parietal Lobule	332	28	-54	60	7.110
	Left Parahippocampal Gyrus	112	-32	-8	-30	6.380
	Right Middle Frontal Gyrus	143	46	16	30	6.244
	Right Postcentral Gyrus	85	58	-18	46	6.026
	Left Middle Frontal Gyrus	134	-28	-4	56	5.897
	Left Insular Cortex	169	-34	-10	14	5.705
NBM (with NMM controlled) > 0						
	Left Occipital Pole	31058	-12	14	46	14.401
	Right Insular Cortex	401	26	32	4	8.876
	Right Superior Parietal Lobule	370	28	-54	60	8.356
	Right Postcentral Gyrus	97	60	-18	44	7.412
	Right Insular Cortex	121	40	-16	14	7.208
	Right Inferior Frontal Gyrus	353	48	16	28	7.008
	Right Parietal Operculum Cortex	140	60	-30	26	6.094
	Left Postcentral Gyrus	231	-48	-34	32	5.659
	Right Precentral Gyrus	108	28	-16	66	5.196
	Right Middle Frontal Gyrus	126	38	2	66	4.984
NBM (with BBM controlled) > 0						
	Right Occipital Pole	8202	-36	-56	-10	10.802
	Left Precentral Gyrus	666	-24	10	26	8.553
	Left Postcentral Gyrus	299	-56	-24	52	8.434
	Left Superior Lateral Occipital Cortex	520	-34	-68	46	8.272
	Left Supplementary Motor Cortex	153	-12	2	50	7.272
	Left Posterior Temporal Fusiform Cortex	252	-34	-20	-34	5.591
	Right Superior Lateral Occipital Cortex	147	44	-82	26	5.294

NBM: Familiar > Novel

Right Temporal Occipital Fusiform Cortex	1469	42	-44	-26	10.48
Left Temporal Fusiform Cortex	543	-42	-46	-28	7.816
Left Inferior Frontal Gyrus	679	-40	18	20	7.222
Left Occipital Pole	118	-28	-102	-6	7.082
Left Occipital Fusiform Gyrus	141	-36	-80	-20	5.682

NBM: Novel > Familiar

Right Superior Lateral Occipital Cortex	1306	38	-66	24	9.995
Left Supramarginal Gyrus	786	-34	-26	32	7.953
Right Supplementary Motor Cortex	97	12	10	56	6.773
Left Superior Lateral Occipital Cortex	522	-30	-68	34	6.488
Left Precentral Gyrus	260	-52	0	38	6.388
Right Precentral Gyrus	684	52	6	30	6.335
Left Lateral Occipital Cortex	269	-28	-78	4	5.976
Left Posterior Inferior Temporal Gyrus	99	-58	-58	-12	4.862

1342 *Note.* All contrasts except 'NBM: Familiar > Novel' and 'NBM: Novel > Familiar' were in learning with familiar objects. Abbreviations: 95CI = 95% confidence interval;

1343 NBM = neural Bayesian model; NMM = neural mean model; BBM = behavioral Bayesian model.

1344

1345 **Table S8**  
1346 *Regions in Harvard-Oxford Atlas showing the advantage of the NBM over the prior-permuted NBM in predicting new word neural representation in learning*  
1347 *with familiar objects (FDR corrected,  $q < .05$ ).*

Region label in HOA	Observed mean	Permuted mean	Permuted 95CI	SES	q
Precuneous Cortex (L)	0.063	0.061	[0.062, 0.061]	4.433	< .001
Occipital Pole (L)	0.249	0.244	[0.246, 0.242]	4.522	< .001
Lateral Occipital Cortex, inferior division (R)	0.355	0.354	[0.355, 0.353]	3.833	.002
Lateral Occipital Cortex, superior division (L)	0.129	0.127	[0.128, 0.126]	3.555	.004
Temporal Occipital Fusiform Cortex (L)	0.157	0.156	[0.157, 0.156]	3.501	.004
Temporal Occipital Fusiform Cortex (R)	0.180	0.179	[0.179, 0.178]	3.235	.009
Occipital Fusiform Gyrus (R)	0.256	0.255	[0.256, 0.254]	3.105	.011

1348 *Note.* Both the observed and permuted values were Fisher-transformed z values. Standard effect size (SES) was calculated as  $\frac{(x-\mu)}{\sigma}$ , where x is the observed mean  
1349 value,  $\mu$  is the mean of the null distribution, and  $\sigma$  is the standard deviation of the null distribution. Abbreviations: 95CI = 95% confidence interval.