LINKING REPRESENTATIONS TO FUNCTION ACROSS NEURAL NETWORK POPULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do differences in internal representations relate to differences in what intelligent systems can do? The growing diversity of trained neural networks creates an opportunity to address this question by treating representations as a data modality with population-level structure of its own. We introduce a metamodeling framework that learns relationships between representational structure and behavior across systems, using representational measurements as predictive markers analogous to biomarkers in medicine. To compare representations across models with different dimensionalities and coordinate systems, we use symmetry-aware descriptions based on invariant geometric summaries and alignment to shared spaces. Across vision and language, representations elicited by shared stimuli predict variation in visual recognition, robustness to distribution shifts, and linguistic capabilities. These relationships generalize to held-out architecture families in vision and model scales in language, revealing regularities in the mapping from representational organization to function. Our analyses further show that accessing this predictive information depends on how representations are measured: natural images provide more informative probes of visual performance, while linguistic predictions depend on the match between probe stimuli and the capability being assessed. Together, these findings establish internal representations as transferable markers of model capabilities and provide a foundation for a comparative science linking representational structure to function across intelligent systems.

1 INTRODUCTION

Recent progress in deep learning research have produced large populations of trained models that vary in architecture, objective, training data and scale (Horwitz et al., 2026; Sellam et al., 2021; Schürholt et al., 2022). This abundance of publicly available trained networks creates an opportunity to study representations at a population level instead of only doing so within individual networks. Each model converts its inputs into representations that preserve some properties of the input while discarding others. The organization of this information influences the behaviors the model can support. Most work on representation analysis of models has focused on understanding how models encode different stimuli (Hewitt & Manning, 2019; Zhang et al., 2020; Tenney et al., 2019) or measuring how strongly representations converge or diverge across different models (Huh et al., 2024; Gröger et al., 2026). This raises a natural question: can differences in internal representations be used to predict variation in behavior across a population of systems? Establishing such relationships would move representational comparison toward identifying structure-function relationships that extend across models.

To address this question, we formulate the relationship between representation and behavior as a population-level meta-modeling problem. Each trained network is one observation, with its representational structure as the predictive input and its behavioral capabilities as the targets. The metamodel learns which patterns of representational variation predict differences in behavior, allowing us to test whether structure-function relationships learned from one population extend to models with different architectures or scales. We also extend the framework to biological systems, testing whether neural representations predict differences in visual decision behavior across mice Angelaki et al. (2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

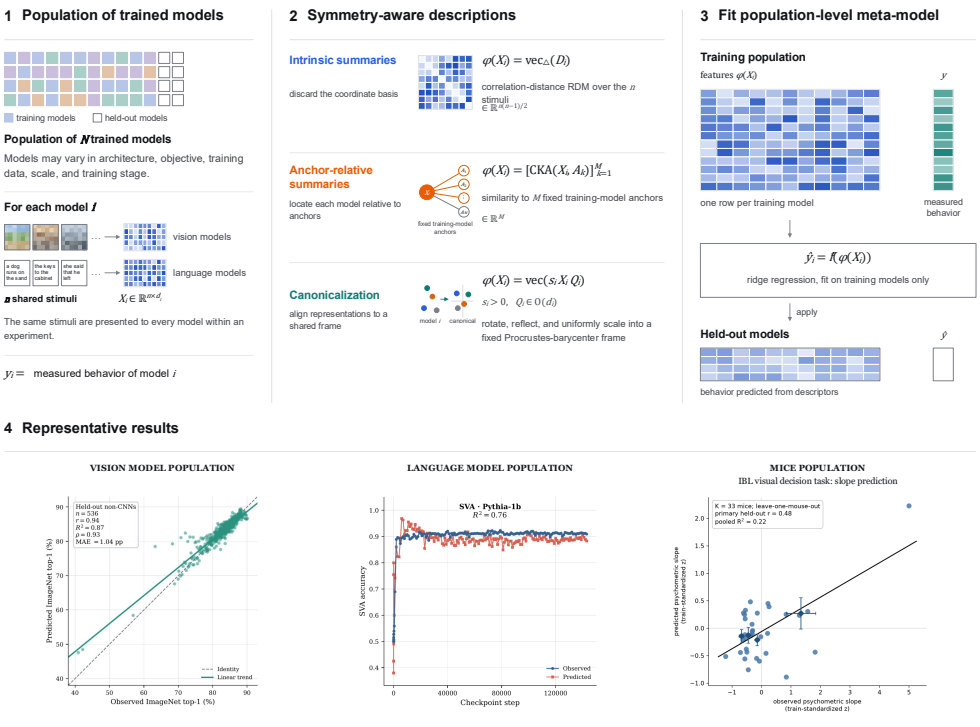


Figure 1: **Symmetry-aware meta-modeling of structure–behavior relationships.** A population of model instances is presented with a shared stimulus set (1). Each model's activations are converted into a comparable description of representational structure: an intrinsic summary, an anchor-relative summary, or canonicalization (2). A meta-model is fit using only the training population (3) and used to predict the behavior of held-out systems (4): ImageNet top-1 accuracy of non-CNN vision models, subject–verb agreement accuracy across training checkpoints of Pythia-1b, and psychometric slope of held-out mice in the IBL visual decision task.

Learning across models requires accounting for symmetries in their internal representations. Permuting hidden units together with their associated connections, for example, changes representational coordinates while leaving the network's input-output function unchanged. A population-level predictor should respect these equivalences while accommodating differences in representational dimensionality across models. We therefore consider two complementary approaches to constructing symmetry-aware descriptions. The first constructs invariant summaries that capture representational structure while remaining unchanged under selected coordinate transformations. These include intrinsic summaries, which describe relationships among stimulus representations within each model, and anchor-relative summaries, which characterize each model through its relationships to a fixed set of reference models. The second approach uses canonicalization to express each model's representations along a shared set of representational axes, allowing the aligned representations to serve directly as predictive inputs.

This formulation lets us ask how broadly structure-function relationships generalize and under what conditions they can be learned. In both vision and language, we examine which stimuli reveal behaviorally informative structure, how many stimuli are needed, and whether representations elicited by the same stimuli can predict different capabilities. We further assess how prediction depends on the choice of symmetry-aware representational description and of meta-model. Finally, to test whether the learned relationships extend beyond the training population, we evaluate on architecture families (vision) and model scales (language) that are held out from training.

We find that across both domains, representational structure can predict behavior with high accuracy in new model populations, suggesting shared structure-function regularities across systems with different designs and sizes. Our results also clarify what is needed to discover these regulari-

ties. Stimuli that support accurate predictions of linguistic abilities, for example, do not necessarily support predictions of commonsense reasoning. The stimuli must reveal distinctions relevant to the capability of interest. Using more stimuli generally improves prediction, while different ways of describing representations prove useful for different populations. These findings pave the way for a new research program that treats variation across intelligent systems as a source of insight into general laws linking how they represent information to what they can do.

2 Related Works

Population-Level Studies of Trained Models: Large collections of trained models have enabled systematic studies of how model properties vary across populations. Horwitz et al. (2026) propose a Model Atlas that represents millions of public models as nodes in a directed graph with edges representing known weight transformations (e.g., fine-tuning). Controlled model zoos provide complementary resources by systematically varying architecture, initialization, and training (Schürholt et al., 2022; Nezhurina et al., 2026; Cherti et al., 2023). In language, MultiBERTs provides independently trained BERT models and Pythia provides checkpoints across model scales and training stages (Sellam et al., 2021; Biderman et al., 2023). These resources make it possible to study how model properties and behaviors vary across populations.

Predicting Model Properties from Neural Network Artifacts: Prior work predicts model properties from several kinds of neural-network artifacts, including weights and learned weight representations (Unterthiner et al., 2020; Schürholt et al., 2021), LoRA weights (Putterman et al., 2024), weight-space networks (Dayan et al., 2026; Shamsian et al., 2024), outputs (Kahana et al., 2025), gradients (Gelberg et al., 2026), and internal activations (Frasca et al., 2026). Related work learns from internal activations to detect hallucinations in individual responses (Frasca et al., 2026). Closer to our setting, representational geometry has been used to predict anatomical hierarchy, network test accuracy, and robustness (Williams et al., 2021; Chou et al., 2026). Building on these directions, we develop a general framework for representational meta-modeling using complementary symmetry-aware descriptions and test how stimuli, behavioral targets, and model populations affect generalization, including to individual differences in mouse behavior.

3 Problem Formulation and Methods

We study a population of N trained model instances. Within an experiment, all models receive the same ordered stimulus set $\mathcal{S} = (s_1, \ldots, s_n)$. Let $X_i \in \mathbb{R}^{n \times d_i}$ denote the response matrix of model i, whose jth row is the response $h_i(s_j)^\top$ to stimulus s_j. The width d_i may vary across models. For vision models, $h_i(s_j)$ is the pooled penultimate-layer activation. For language models, it is the activation of the last non-padding token in the penultimate transformer block.

Each model has a behavioral score $y_i^{(t)}$ on task t. These targets include ImageNet top-1 accuracy and mean out-of-distribution accuracy in vision, and linguistic and commonsense reasoning accuracy in language. For a training population $\mathcal{T}$ and a disjoint evaluation population $\mathcal{E}$, we construct a fixed-length descriptor and fit a separate predictor for each target,

$$v_i = \phi_{\mathcal{T}}(X_i) \in \mathbb{R}^{q_\phi}, \qquad \widehat{y}_i^{(t)} = f_{t,\mathcal{T}}(v_i). \tag{1}$$

Here $f_{t,\mathcal{T}} \colon \mathbb{R}^{q_\phi} \to \mathbb{R}$ is the meta-model that maps a model's representational descriptor to its predicted behavioral score on task t. The descriptor dimension q_ϕ is shared across models within an experiment. The subscript $\mathcal{T}$ allows the descriptor to depend on reference models or alignment parameters estimated from the training population.

3.1 Symmetry-aware representational descriptions

We propose descriptions for this meta-modeling problem using two complementary constructions. The first constructs invariant summaries from either pairwise dissimilarities among a model's representations of the shared stimuli (intrinsic summaries) or comparisons with the representations of fixed reference models (anchor-relative summaries). The second uses canonicalization to align representations along shared representational axes. We define each description below.

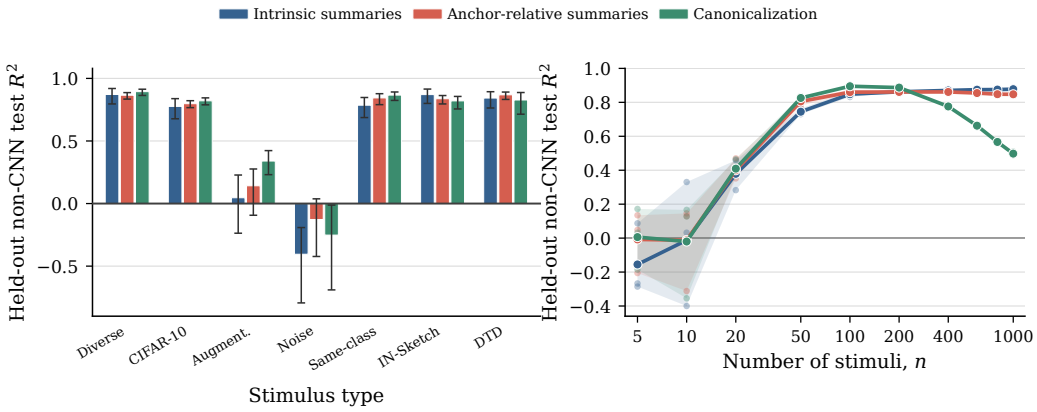


Figure 2: **Effects of stimulus type and count on IID prediction.** Held-out non-CNN R^2 for ridge meta-models using different summary representation methods **(a)** across stimulus types and **(b)** across numbers of diverse ImageNet stimuli. Error bars in (a) show 95% bootstrap CIs over held-out models; shaded regions in (b) show the range across three stimulus subsets.

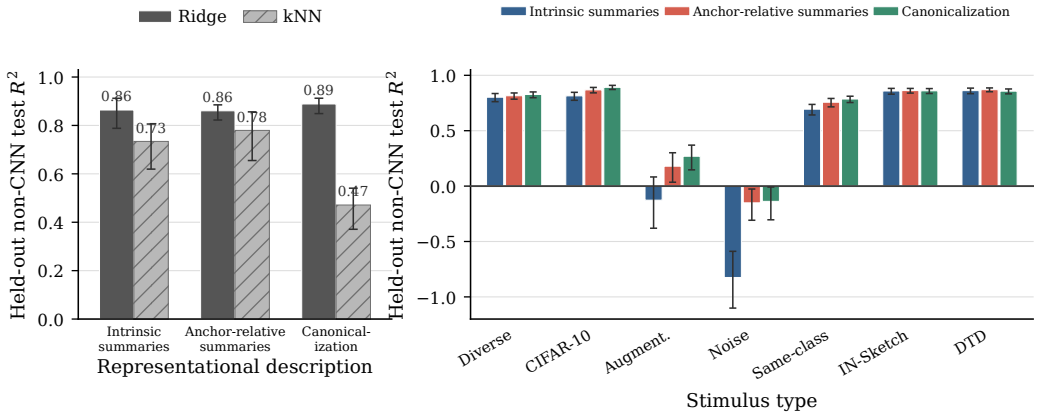


Figure 3: **Prediction method and OOD behavior.** **(a)** Ridge and kNN prediction of ImageNet accuracy using three representational descriptions of the same 200 diverse stimuli. **(b)** Ridge prediction of mean OOD accuracy across stimulus conditions. Error bars show 95% paired-bootstrap confidence intervals over held-out non-CNN models.

Intrinsic summaries. We describe each model using a representational dissimilarity matrix (RDM) with Pearson correlation distance (Kriegeskorte et al., 2008). To make its invariances explicit, let $C_d = I_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^\top$ and define

$$u_{ij} = \frac{C_{d_i} h_i(s_j)}{\|C_{d_i} h_i(s_j)\|_2}, \qquad U_i = [u_{i1}, \ldots, u_{in}]^\top, \tag{2}$$

This operation centers each stimulus response across units and normalizes its length. The RDM satisfies

$$\begin{aligned} D_i(j,k) &= 1 - \mathrm{corr}\big(h_i(s_j), h_i(s_k)\big) \\ &= 1 - u_{ij}^\top u_{ik}, \qquad D_i = \mathbf{1}_n\mathbf{1}_n^\top - U_i U_i^\top. \end{aligned} \tag{3}$$

We use its strict upper triangle, in a fixed order, as $\phi_{\mathrm{int}}(X_i) \in \mathbb{R}^{n(n-1)/2}$. This descriptor has the same dimension for every model and requires no fitted reference population. The summary is invariant to permutation of units, positive rescaling or a uniform offset of each stimulus response, and orthogonal transformations that preserve the all-ones direction, satisfying $Q\mathbf{1}_{d_i} = \pm\mathbf{1}_{d_i}$. More gen-

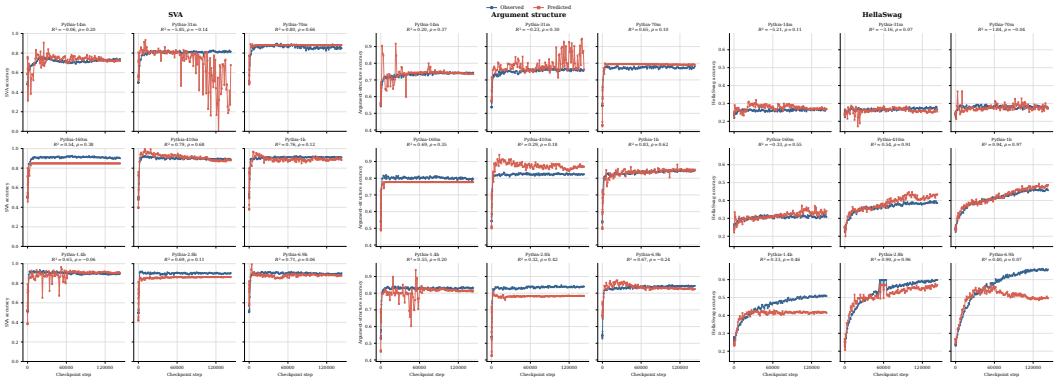


Figure 4: **Observed and predicted behavioral trajectories across language-model checkpoints.** Each group shows leave-one-model-size-out predictions for SVA, argument structure, or HellaSwag using ridge regression on RDM features constructed from 200 stimuli. Blue curves show observed accuracy and red curves show predicted accuracy. Panel titles report R^2 and Spearman ρ.

erally, the following completeness result shows that the RDM determines the standardized responses up to an arbitrary orthogonal transformation.

Proposition 3.1 (Completeness of Gram and RDM summaries) *For $U, V \in \mathbb{R}^{n\times p}$,*

$$UU^\top = VV^\top \quad \Longleftrightarrow \quad V = UQ \text{ for some } Q \in O(p).$$

Consequently, two models have identical correlation RDMs if and only if their standardized response matrices, zero-padded to a common width, are related by an orthogonal transformation.

Suppose $UU^\top = VV^\top$. The map $U^\top a \mapsto V^\top a$ is well-defined and preserves inner products, since

$$\langle U^\top a, U^\top b\rangle = a^\top UU^\top b = a^\top VV^\top b = \langle V^\top a, V^\top b\rangle.$$

It therefore extends to an orthogonal map on $\mathbb{R}^p$, giving $V = UQ$ for some $Q \in O(p)$. The converse follows from $QQ^\top = I_p$. The RDM statement follows from $D_i = \mathbf{1}_n\mathbf{1}_n^\top - U_iU_i^\top$ and the fact that zero-padding preserves Gram matrices.

Anchor-relative summaries. We also construct invariant summaries relative to an ordered set of training anchors $\mathcal{A}_\mathcal{T} = (a_1, \ldots, a_M)$. We use all training models as anchors, so $M = |\mathcal{T}|$, and define $\phi_{\text{anc},\mathcal{T}}(X_i) = \left[\text{CKA}(X_i, X_{a_1}), \ldots, \text{CKA}(X_i, X_{a_M})\right]^\top$, where CKA stands for linear centered kernel alignment (CKA) (Kornblith et al., 2019). Let $H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ and $K_i = H_nX_iX_i^\top H_n$. Then $\text{CKA}(X_i, X_a) = \frac{\langle K_i, K_a\rangle_F}{\|K_i\|_F\|K_a\|_F}$, for nonzero K_i and K_a. Here centering is across stimuli, in contrast to the centering across units in equation 2. Each score, and hence the anchor-relative descriptor, is invariant to independent orthogonal transformations and global nonzero rescalings of the model and anchor representations. Its dimension is independent of their widths. Held-out models use the same fixed training anchors.

Canonicalization. To use activation patterns directly as predictive inputs, we first express representations in a common basis. Treating rotations and reflections of representational axes as nuisance variation, we estimate a Procrustes barycenter as a shared reference to which all models can be aligned. The barycenter is learned from the training population and then held fixed when aligning new models (Gower, 1975). Let $Z_i \in \mathbb{R}^{n\times p}$ denote X_i after zero-padding to a common width. We estimate the barycenter by approximately minimizing

$$\min_{\|B\|_F=1,\ s_i\geq 0,\ R_i\in O(p),\ i\in\mathcal{T}} \sum_{i\in\mathcal{T}} \|s_iZ_iR_i - B\|_F^2, \tag{4}$$

where R_i aligns representational axes and s_i accounts for global scale. We alternate between aligning each training representation to the current template and updating the template to the normalized

mean of the aligned representations. The fitted barycenter $\widehat{B}_{\mathcal{T}}$ defines the shared basis. Each training or held-out model is aligned to it by solving $(\widehat{s}_i, \widehat{R}_i) \in \arg\min_{s\geq 0 \mathbb{R}\in O(p)} \|sZ_iR - \widehat{B}_{\mathcal{T}}\|_F^2$. We then use the aligned activations as predictive inputs: $\phi_{\mathrm{can},\mathcal{T}}(X_i) = \mathrm{vec}(\widehat{s}_i Z_i \widehat{R}_i) \in \mathbb{R}^{np}$.

Meta-model fitting and evaluation. We focus primarily on ridge regression, asking whether behavioral differences across models can be predicted from a linear combination of representational features. For each behavioral target, we fit a separate ridge model (unpenalized intercept) on standardized descriptors of the training models, selecting the regularization strength λ by cross-validation within the training population. As a complementary analysis, we evaluate k-nearest-neighbor (kNN) regression, which predicts behavior from the scores of nearby training models in descriptor space. This tests whether local relationships among models also support generalizable predictions. To test whether the learned relationships generalize across model designs and sizes, we evaluate on held-out architecture families in vision and held-out model scales in language. All checkpoints of each held-out language model scale are excluded from training. We report R^2 on each held-out population to quantify how well the predictions account for variation in behavioral scores.

4 Vision Results

Experimental Setup: We test generalization across architectures by training meta-models on CNNs and evaluating their predictions on held-out non-CNN models. Our primary analysis includes 1,259 pretrained image-classification networks from `timm` (Wightman, 2019), comprising 723 CNNs and 536 non-CNN models. We predict ImageNet top-1 accuracy and accuracy on 17 out-of-distribution datasets from the Model-vs-Human benchmark (Geirhos et al., 2021). These targets allow us to ask whether representational structure predicts both performance on the training distribution and robustness to distribution shifts. We examine how prediction depends on both the number and type of stimuli used to elicit representations. To vary stimulus number, we sample diverse ImageNet images with $n \in \{5, 10, 20, 50, 100, 200, 400, 600, 800, 1000\}$. To vary stimulus type, we compare diverse ImageNet images (Russakovsky et al., 2015), CIFAR-10 images (Krizhevsky et al., 2009), stochastic augmentations of a single ImageNet image, synthetic noise, images from a single ImageNet class, ImageNet-Sketch images (Wang et al., 2019), and DTD textures (Cimpoi et al., 2014).

4.1 Results

Structure-function relationships generalize across architectures. We first ask whether a structure-function relationship learned on one architecture family generalizes over to models of different family. Using representations extracted by only 200 diverse ImageNet images, ridge meta-models trained on CNNs predict the ImageNet accuracy of held-out non-CNN architectures with $R^2 = 0.86$, 0.86, and 0.89 for the invariant, anchor-relative, and canonicalization-based summaries, respectively (Figure 1(4)left, 2a). This shows that representations of a small shared stimulus set predict performance across architecture families without evaluating models on the target task. Thus, much of the variation in performance across unseen architecture families can be recovered from how models internally organize a small common set of inputs.

Which stimuli reveal structure-function? We then ask how many stimuli, and of what kind, are needed for strong predictions. We find that with 10 or fewer stimuli, prediction performance is near chance. Performance rises sharply between 20 and 50 images and reaches $R^2 \approx 0.85$ by $n = 100$ (Figure 2b). The intrinsic and anchor-relative summaries stay at this level up to $n = 1000$. The canonicalization-based approach peaks at $n = 100$–200 and then declines. Interestingly, we find that broad coverage of ImageNet is not required for good prediction performance. CIFAR-10, exemplars from a single ImageNet class, ImageNet-Sketch, and DTD textures all give strong R^2 values. Critically, we find that inputs that do not have any natural image structure behave very differently. Two hundred augmentations of a single image give at most $R^2 = 0.34$, and just synthetic noise gives negative R^2 for all three summaries. Thus, natural image regularities are need for learning generalizable structure-function relationships.

In-distribution representations predict OOD behavior. Representations of standard natural images also help us predict how held-out non-CNN models behave under distribution shift. With just 200 diverse ImageNet images, ridge meta-models predict mean accuracy across the 17 Model-vs-Human OOD datasets with $R^2 = 0.80$–0.83 across our three summary descriptions. (Figure 3b).

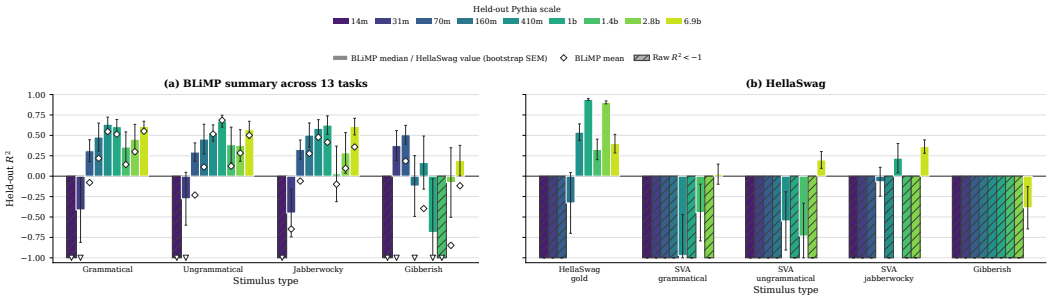


Figure 5: **Effect of stimulus content on cross-scale language prediction.** Each ridge meta-model uses intrinsic-summary features constructed from 200 stimuli and is evaluated on checkpoints from a held-out Pythia scale. **(a)** Bars show median R^2 across 13 BLiMP task groups; hollow diamonds show the corresponding mean. **(b)** Bars show HellaSwag R^2. Hatched bars and downward triangles indicate raw $R^2 < -1$, displayed at -1. Error bars show standard errors over held-out checkpoints.

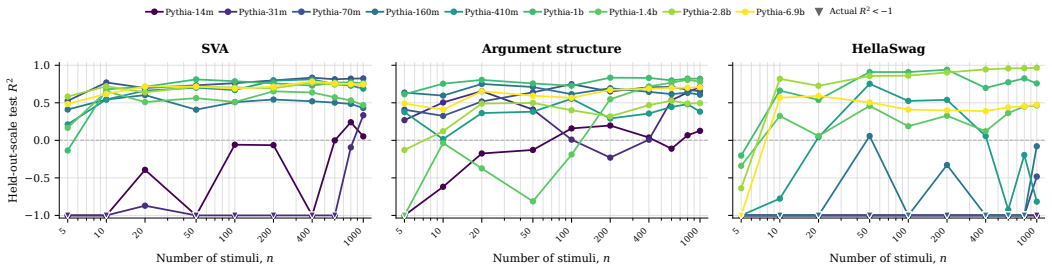


Figure 6: **Effect of stimulus count on language-model prediction.** Each panel shows ridge prediction from invariant summary features: SVA, argument structure, or HellaSwag. The number of stimuli used to construct varies from 5 to 1,000. Displayed R^2 values are clipped to $[-1, 1]$; downward triangles indicate values below -1.

Here, the relation between performance and stimulus type also closely follows the IID results. This highlights that differences in how models represent a small set of normal images carry information about robustness even though none of these stimuli are extracted from the OOD benchmark datasets.

A learned mapping outperforms nearest-neighbor prediction. We also use k-nearest-neighbor regression, which predicts the same target behvaior as the average of its three most similar training models as a non parametric baseline. The motivation behind this baseline is to test whether prediction can be explained by representational similarity alone. Ridge regression outperforms k-NN for all three summaries (Figure 3a). This consistent advantage of ridge suggests that prediction relies on a relationship learned across the whole training population, rather than on the similarity alone.

5 Language Results

Experimental Setup: We test whether structure-function relationships learned across language models generalize to models of a different size. Our population comprises 1,383 checkpoints from nine Pythia scales, ranging from 14M to 6.9B parameters (Biderman et al., 2023). Each checkpoint is treated as a separate model instance. The behavioral targets comprise 13 linguistic task groups derived from BLiMP (Warstadt et al., 2020) and commonsense sentence-completion accuracy on HellaSwag (Zellers et al., 2019). To examine how prediction depends on stimulus number, we vary the number of sentences over $n \in \{5, 10, 20, 50, 100, 200, 400, 600, 800, 1000\}$. To examine stimulus content, we fix $n = 200$. For BLiMP, we compare grammatical sentences, their paired ungrammatical alternatives, jabberwocky sentences that retain the grammatical template while replacing content words with pseudowords, and gibberish sequences assembled from random words. For HellaSwag, we use contexts concatenated with their correct endings. We also predict HellaSwag performance

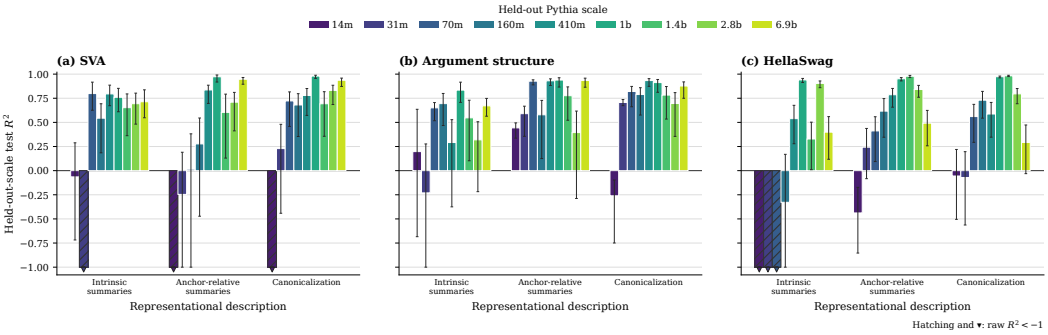


Figure 7: **Comparison of representational descriptions across language tasks.** Error bars show 95% paired-bootstrap confidence intervals over held-out checkpoints.

using the four stimulus types used for the BLiMP subject-verb agreement (SVA) task (Appendix C.1), testing whether stimuli informative about linguistic competence also support predictions of commonsense reasoning. In each of nine folds, the meta-model is trained on checkpoints from eight scales and evaluated on all checkpoints from the remaining scale. This tests whether representations predict a model's behavioral trajectory across training even when its scale is absent from the training population. We treat ridge regression on intrinsic summaries as the primary meta-model, and we report R^2 within each held-out scale.

5.1 Results

Representational structure predicts behavior across model scales. We ask the same question about the structure-function relationship in the language setting. Using representations of 200 stimuli, ridge meta-models trained on eight scales recover the learning trajectory of the held-out scale (Figure 4). Test R^2 is positive for 7 of 9 held-out scales on SVA, 8 of 9 on argument structure, and 5 of 9 on HellaSwag. The best folds reach $R^2 = 0.80$ for SVA at 70M, 0.83 for argument structure at 1B, and 0.94 and 0.90 for HellaSwag at 1B and 2.8B, respectively. For SVA, we can see that the predictions follow the early rise in accuracy from chance to about 0.9, and at larger scales, R^2 stays between 0.65 and 0.79. The negative R^2 values for HellaSwag at 14M–160M are not caused by large errors, but rather by accuracy remaining near chance throughout training at these scales, such that even small errors of 1.3%–1.7% points are large relative to the variation that R^2 measures.

Stimulus content determines how well representations predict function. Figure 5 shows how the choice of stimuli affects the predictive power of the meta-model. For BLiMP tasks, we aggregate held-out R^2 over the 13 task groups. We find that grammatical, ungrammatical, and jabberwocky sentences behave similarly. This suggests that this retained structure of jabberwocky sentences helps expose behaviorally informative differences between models. Gibberish sentences are substantially less informative, with mean R^2 across task groups below zero at eight of nine held-out scales. This is analogous to the vision results, where noise images and augmented images were far less informative than natural images. HellaSwag shows a much stronger dependence on stimulus content. The same linguistic stimuli do not support accurate predictions of commonsense reasoning with this approach, whereas HellaSwag contexts paired with their correct endings support positive R^2 at every scale from 410M onward. Thus, the ability to predict behavior from representations depends on which aspects of a model's capabilities the stimuli engage. This shows that stimulus selection is an important part of the scientific problem of discovering structure-function relationships.

A small set of sentences can reveal substantial differences in capability Figure 6 varies the number of stimuli used to build the invariant summary from 5 to 1,000. We see that much of the predictive information is already present in responses to a small shared stimulus set. For the SVA task, we find that even a small stimulus set yields positive R^2: the median held-out R^2 across scales rises from 0.21 at $n = 5$ to 0.62 at $n = 10$ and stabilizes from $n = 50$ onward. However, the two smallest models (14M and 31M) remain negative until n reaches 800-1,000. For Argument Structure task, its median R^2 rises gradually from $n = 5$ to about 0.65 at $n \geq 400$. HellaSwag performance at 1B, 2.8B, and 6.9B is also consistently well predicted from ten stimuli onward with the 2.8B model pre-

diction reaching $R^2 = 0.97$ at $n = 1{,}000$. These trends are also consistent with the vision setting, where R^2 rises between 20 and 100 stimuli and changes little beyond 200. In both domains, a few hundred stimuli are enough to capture most of the predictive signal.

The intrinsic summary falls behind on harder tasks. The generality of a learned structure-function relationship also depends on the representational descriptions (Figure 7). On SVA, all three approaches perform similarly. For argument structure, the anchor-relative summary and canonicalization reach median R^2 values of 0.78 and 0.79, respectively, compared with 0.55 for the intrinsic summary. HellaSwag shows the largest gap in performance between the intrinsic summary and the other two methods. The intrinsic summary is positive for only five of the nine held-out scales and falls below $R^2 = -1$ for the three smallest models, whereas the anchor-relative summary and canonicalization are positive for eight and seven scales. In vision, the three descriptions perform almost identically, whereas in language this choice matters more for harder tasks like HellaSwag.

6 Brain Results

We next test whether metamodeling extends from populations of artificial networks to biological individuals, i.e. can differences in neural representational geometry predict differences in behavioral sensitivity? We test this by analyzing 33 mice performing the International Brain Laboratory visual contrast-discrimination task (Angelaki et al., 2025) (Appendix Figure 24A). For each mouse, we use separate halves of the trials to construct neural representations and estimate behavioral sensitivity, measured by the slope of a psychometric curve (Appendix Figure 24B)). Trial-averaged responses from 100 neurons across nine signed contrasts and ten time bins form an activation matrix, from which we compute an RDM using pairwise Euclidean distances. We use the vectorized RDM as input to ridge regression and evaluate psychometric-slope prediction with leave-one-mouse-out cross-validation across five trial splits. Full experimental details are provided in Appendix D.1.

Results. Neural representational geometry predicts behavioral sensitivity in held-out mice, achieving mean pooled z-$R^2 = 0.22$ and Pearson's $r = 0.48$ across trial splits (last panel of Figure 1; per-mouse predictions in Appendix Figure 24C). In mice, intrinsic summaries outperform anchor-relative and canonicalized representations (Appendix Figure 24D). These results support the broader premise that representational geometry captures behaviorally relevant differences across systems, even when those systems are individual brains with no direct correspondence between their recorded neurons. In the appendix D.2, we report results for temporal-resolution scaling and contrast subsets.

7 Discussion

We showed that differences in how neural networks represent a shared set of inputs predict differences in what they can do, and that these relationships generalize beyond the population used to learn them. Across vision and language, meta-models fit on one population of models predicted recognition performance, robustness to distribution shift, and linguistic and reasoning abilities in populations held out by architecture family or scale. The same approach extended to biological systems, where neural representational geometry predicted individual differences in perceptual sensitivity across mice.

A growing body of work asks whether different neural networks, and even brains, converge on similar representations of the world (Huh et al., 2024; Yamins et al., 2014). Our work complements this focus on universality or convergence by asking what can be learned representational divergence. This mirrors a long tradition in comparative anatomy, where differences across species, such as variation in beak shape among birds with different diets, have been a key source of insight into how structure supports function. Linking structure to function within a single system is difficult, but patterns that hold across many systems can reveal which structural differences matter. Our results apply the same logic applies to the growing variety of trained models where differences in how models represent a shared set of inputs track differences in what they can do. This suggests that representations can serve as signs of a model's abilities, much as biomarkers do in medicine. Like biomarkers, these signs are first identified through their association with behavior and can then be tested to see whether they actually drive it. And because the same approach works for artificial networks and animal brains, it offers a shared way to study why individual models, animals, and perhaps people differ in what they can do.

LLM usage disclosure: AI tools were used for minor writing, formatting, and visual editing assistance for generating illustrative figures. The authors take full responsibility for the final content of the paper

REFERENCES

Dora Angelaki, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Kcénia Bougrova, Sebastian A Bruijns, Matteo Carandini, Joana A Catarino, et al. A brain-wide map of neural activity during complex behaviour. *Nature*, 645(8079):177–191, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. Diagnosing generalization failures from representational geometry markers. *arXiv preprint arXiv:2603.01879*, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adir Dayan, Yam Eitan, and Haggai Maron. On the expressive power of permutation-equivariant weight-space networks. *arXiv preprint arXiv:2602.01083*, 2026.

Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Haggai Maron. Neural message-passing on attention graphs for hallucination detection. In *International Conference on Learning Representations*, volume 2026, pp. 449–475, 2026.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Yoav Gelberg, Yam Eitan, Aviv Navon, Aviv Shamsian, Theo Putterman, Michael Bronstein, and Haggai Maron. Gradmetanet: An equivariant architecture for learning on gradients. *Advances in Neural Information Processing Systems*, 38:154115–154158, 2026.

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. We should chart an atlas of all the world's models. *Advances in Neural Information Processing Systems*, 38, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Jonathan Kahana, Eliahu Horwitz, Imri Shuval, and Yedid Hoshen. Deep linear probe generators for weight space learning. In *International Conference on Learning Representations*, volume 2025, pp. 16711–16728, 2025.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024.

Marianna Nezhurina, Tomer Porian, Giovanni Puccetti, Tommie Kerssies, Romain Beaumont, Mehdi Cherti, and Jenia Jitsev. Scaling laws for robust comparison of open foundation language-vision models and datasets. *Advances in Neural Information Processing Systems*, 38:153979–154018, 2026.

Theo Putterman, Derek Lim, Yoav Gelberg, Stefanie Jegelka, and Haggai Maron. Learning on loras: Gl-equivariant processing of low-rank weight spaces for large finetuned models. *arXiv preprint arXiv:2410.04207*, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Konstantin Schürholt, Dimche Kostadinov, and Damian Borth. Self-supervised representation learning on neural network weights for model characteristic prediction. *Advances in Neural Information Processing Systems*, 34:16481–16493, 2021.

Konstantin Schürholt, Diyar Taskiran, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Model zoos: A dataset of diverse populations of neural network models. *Advances in Neural Information Processing Systems*, 35:38134–38148, 2022.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*, 2021.

Aviv Shamsian, Aviv Navon, David W Zhang, Yan Zhang, Ethan Fetaya, Gal Chechik, and Haggai Maron. Improved generalization of weight space networks via augmentations. *arXiv preprint arXiv:2402.04081*, 2024.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. *arXiv preprint arXiv:2002.11448*, 2020.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Daniel LK Yamins et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. In *PNAS*, 2014.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 292–299, 2020.

A Limitations and Future Work

The approach also has important limitations. Most importantly, it shows which features of representations go along with good or poor behavior (i.e., its predictions are correlational), but not whether those representational signatures are responsible for it, and factors such as model size or training data could shape both. Nonetheless, we see several promising future directions for this meta-modeling approach. First, rather than passively comparing stimulus sets, one could optimize or synthesize stimuli to maximize predictive performance, which may reveal which input distinctions most sharply separate more and less capable systems. Second, the framework could also be extended to more diverse behaviors, such as calibration, adversarial robustness, human-like error patterns, and safety-relevant tendencies. Third, it could also be combined with interventions on representations to move from predictive markers toward causal accounts. And fourth, although we used CKA to construct anchor-relative descriptions, any of the many available representational similarity measures (Klabunde et al., 2025) could take its place, so the framework also offers a behaviorally grounded way to compare these measures by how much function-relevant information each preserves. Together, these directions could help establish a comparative science that uses variation across intelligent systems to uncover general principles relating representation to function.

B Additional Vision Experiments

B.1 RDM Metric Variants

The main experiments in the paper use correlation-distance RDMs for intrinsic summary computations. We also compared min-max normalized correlation, min-max normalized Euclidean, and cosine RDMs, and a geotopological RDM (Lin & Kriegeskorte, 2024) obtained by rank-normalizing Euclidean distances and applying a $\mathrm{GT}(0.05, 0.90)$ transform (Figure 8) to vision tasks. On diverse ImageNet $n = 200$ stimuli, all five variants yield similar held-out non-CNN R^2. We therefore keep the unnormalized correlation RDM as the default.

B.2 OOD Scaling with Stimulus Count

Figure 9 shows OOD prediction as the number of diverse ImageNet stimuli increases. Ridge models trained with intrinsic summaries improve rapidly up to approximately $n = 200$ and then saturate, showing the same scaling trend as in the other scaling experiments. Thus, the same small probe set used in the main text is also sufficient for OOD prediction.

B.3 Full OOD Prediction Heatmaps

The main text reports aggregate OOD prediction. Figures 10 and 11 show the corresponding 7×20 grids: seven stimulus conditions by the 17 OOD datasets plus overall, parametric, and nonparametric aggregate versions of the dataset. Ridge and kNN give the same qualitative pattern across all three summaries. Diverse, CIFAR, same-class, ImageNet-Sketch, and DTD stimuli predict held-out non-CNN OOD accuracy well; augmented and noise stimuli do not.

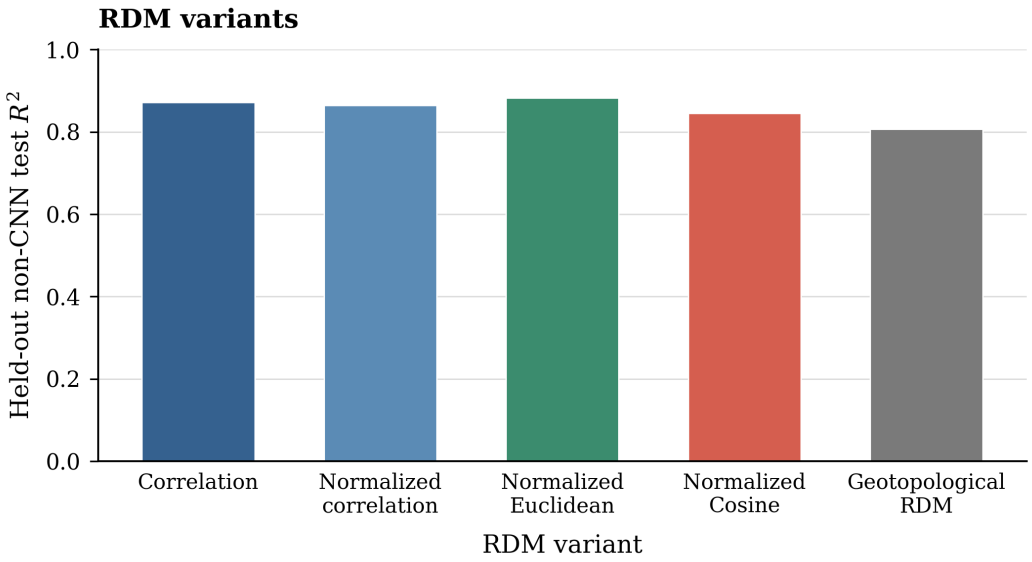


Figure 8: **RDM variants.** Held-out non-CNN R^2 for ridge prediction of ImageNet top-1 accuracy from diverse ImageNet $n = 200$ RDM vectors. Correlation is the paper default. The normalized version uses min-max scaling; the geotopological RDM is a rank-normalized Euclidean RDM with a GT$(0.05, 0.90)$ transform.

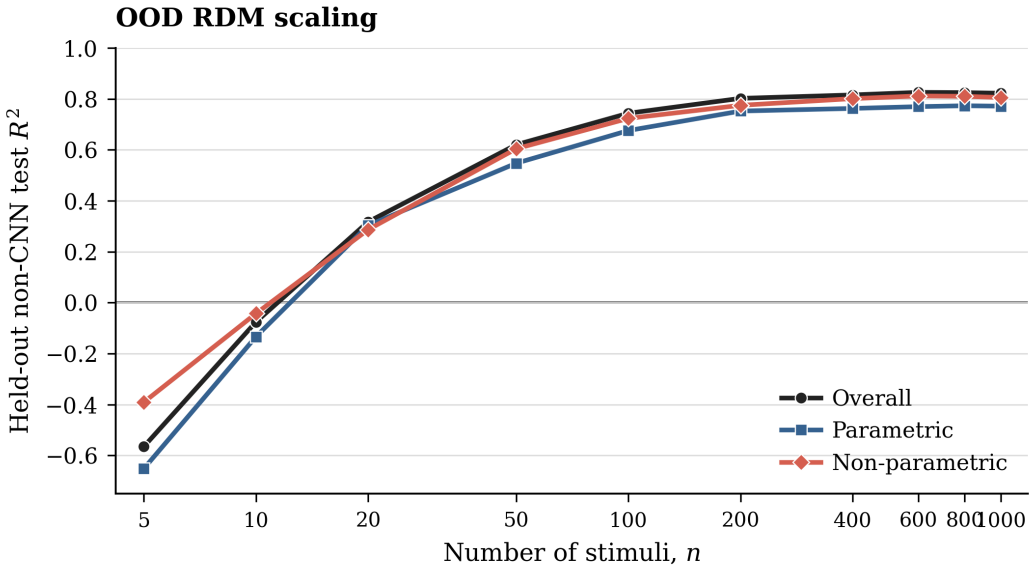


Figure 9: **OOD RDM scaling.** Held-out non-CNN R^2 for correlation-RDM ridge models predicting mean OOD accuracy from diverse ImageNet stimuli. Overall is the mean over 17 OOD datasets; parametric and non-parametric are the corresponding subset means.

B.4 Additional kNN Experiments for Vision

For our main vision experiments, we have set $k = 3$ using Spearman distance for the intrinsic and anchor-relative summaries and cosine distance for the canonicalization-based summary. We swept $k \in \{2, \ldots, 10\}$ and compared Euclidean, cosine, correlation, and Spearman distances on diverse ImageNet $n = 200$ for instrinsic summaries(Figure 12).

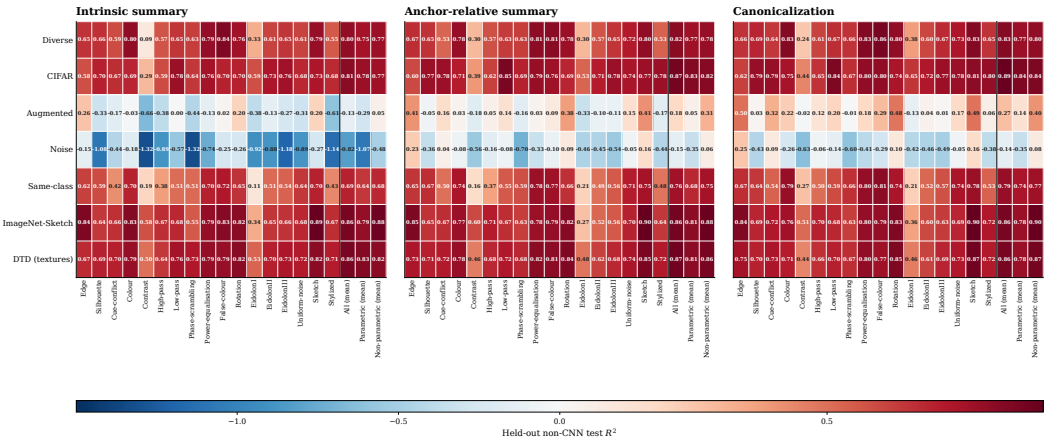


Figure 10: **Ridge OOD heatmaps.** Held-out non-CNN R^2 for ridge meta-models using our proposed three summaries. Rows are stimulus conditions; columns are the 17 Model-vs-Human datasets plus three aggregates.

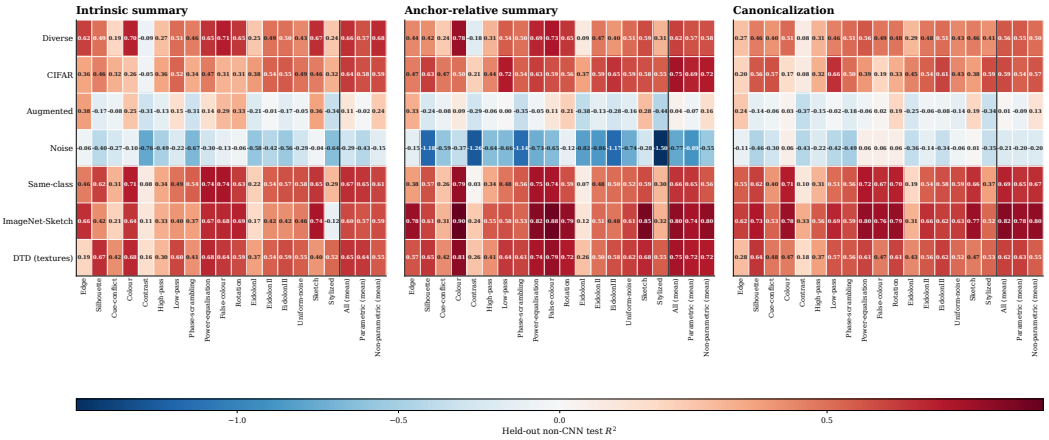


Figure 11: **kNN OOD heatmaps.** Same layout as Figure 10, but for kNN.

B.5 kNN Stimulus-Count Scaling

Figure 13 shows the kNN scaling results on the same paired stimulus subsets as the ridge scaling experiment. Instrinsic and Anchor-relative summaries improve rapidly with n and plateau by roughly $n = 200$–400; Anchor-relative summaries reach the highest held-out non-CNN R^2 at large n. The Canonicalization kNN curve rises more slowly and remains below the other summaries.

C Additional Langauge Results

C.1 Example Stimuli for Language Tasks

The following examples are drawn from the stimulus sets used in our experiments. The grammatical, ungrammatical, and jabberwocky examples are derived from the same SVA minimal pair:

- **(1) Grammatical:** A cousin of a lot of dancers anticipates Naomi to aggravate Angela.
- **(2) Ungrammatical:** A cousin of a lot of dancers anticipate Naomi to aggravate Angela.
- **(3) Jabberwocky:** A coakin of a lon of doffers antecimists Naomi to adquapate Angela.

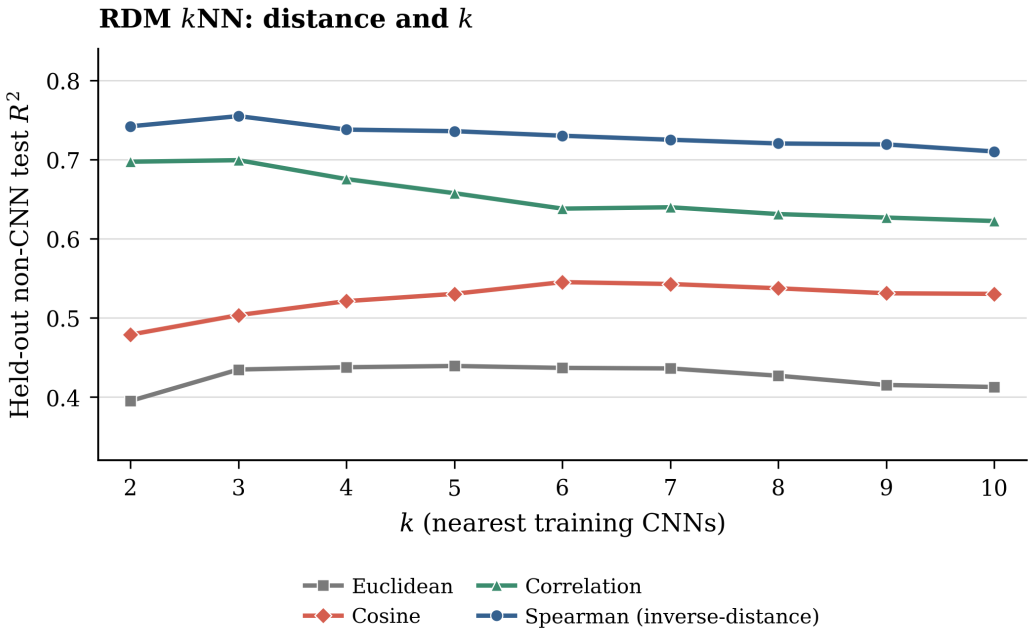


Figure 12: ***k*NN sensitivity to k and distance.** Held-out non-CNN R^2 for predicting ImageNet top-1 accuracy from diverse ImageNet $n = 200$ with instrinsic representation summaries.

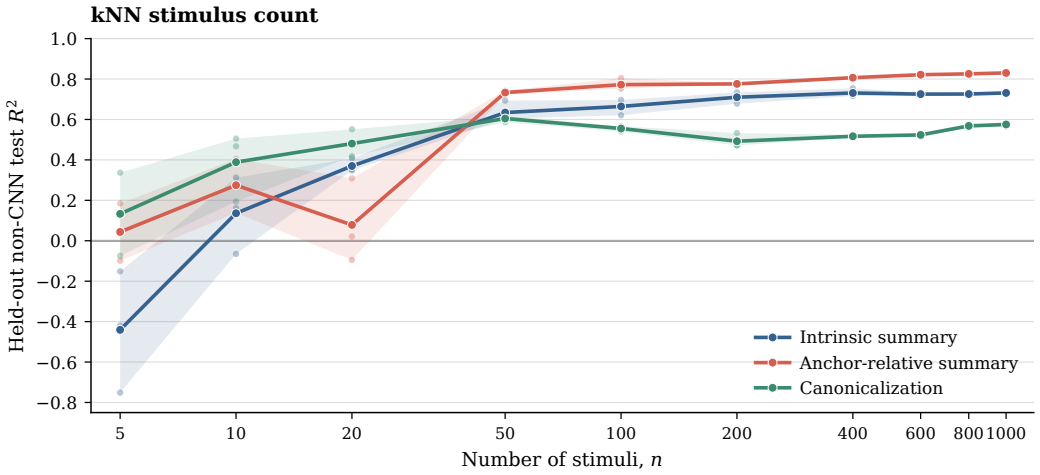


Figure 13: **kNN stimulus-count scaling.** Held-out non-CNN R^2 for kNN meta-models using different summaries of diverse ImageNet stimuli. Curves show the mean across three random subsets; shaded regions show the minimum-to-maximum range, and faint points show individual subsets.

- **(4) Gibberish:** Retrieved movement au execution fruits outdoors initially cup revelation chemical soldier marsh.

- **(5) HellaSwag Context + Gold ending:** A woman pours ice into a glass. She puts shots of alcohol into the glass. She shakes it and pours it into another glass. She puts a straw in it and holds it up.

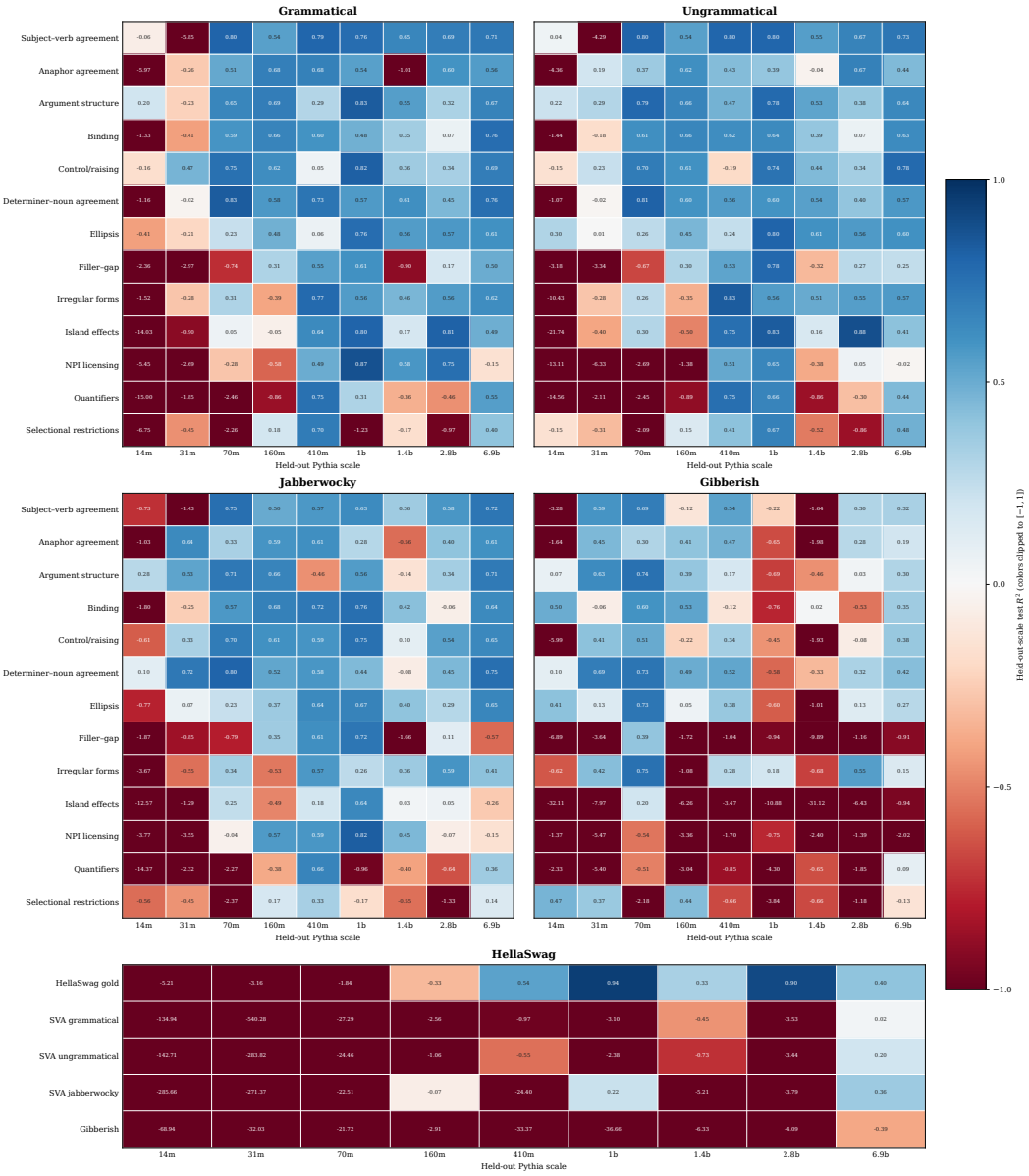


Figure 14: **Per-scale prediction across stimulus types.** Held-out-scale test R^2 for ridge meta-models using intrinsic summaries of 200 stimuli. Each column is one held-out Pythia scale. The top four panels show the 13 BLiMP targets for each stimulus type; the bottom panel shows HellaSwag. Colors are clipped to $[-1, 1]$; annotations report the actual R^2.

C.2 Per-Scale Prediction across Stimulus Types

Figure 14 reports the per-fold values for the aggregates in Figure 5. Each cell gives the held-out R^2 of a ridge meta-model trained on intrinsic summaries from eight Pythia scales and evaluated on the held-out model. Across the 13 linguistic targets, grammatical, ungrammatical, and jabberwocky stimuli behave similarly. Gibberish is positive in only 53 cells, has a median of -0.22, and falls below -1 in 36 cells, roughly twice as many as for any other stimulus type.

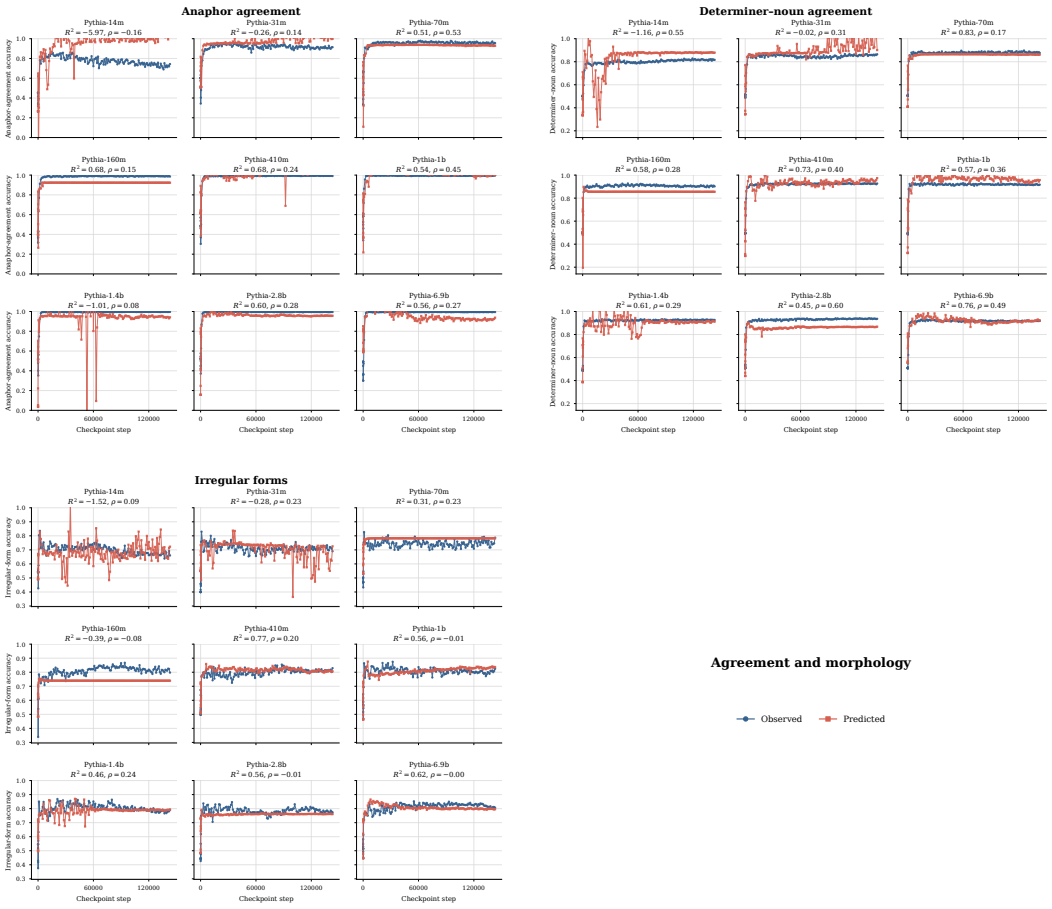


Figure 15: **Agreement and morphology trajectories.** Observed and predicted trajectories for anaphor agreement, determiner–noun agreement, and irregular forms.

C.3 Checkpoint trajectories for the remaining BLiMP tasks

Figures 15–17 show the observed and predicted checkpoint trajectories for the remaining BLiMP tasks. Each task contains nine leave-one-scale-out panels based on intrinsic summaries of 200 grammatical stimuli.

C.4 Representational descriptions across all language tasks

Figures 18 and 19 provide the per-scale results for all language tasks. They complement the results in the main text (Figure 7).

C.5 Stimulus-count scaling for the remaining BLiMP tasks

Figures 20–22 show stimulus-count scaling for the remaining BLiMP tasks. Each curve reports held-out-scale R^2 using intrinsic summaries, complementing the representative tasks shown in the main text.

C.6 Ridge vs kNN in the Language setup

Ridge and kNN succeed on different model scales. Figure 23 compares ridge regression with a fixed kNN predictor using the same RDM features and LOMO splits. The kNN predictor improves six of nine SVA folds and seven of nine folds for both argument structure and HellaSwag. Its largest gains occur where ridge extrapolation fails: for example, HellaSwag improves from $R^2 =$

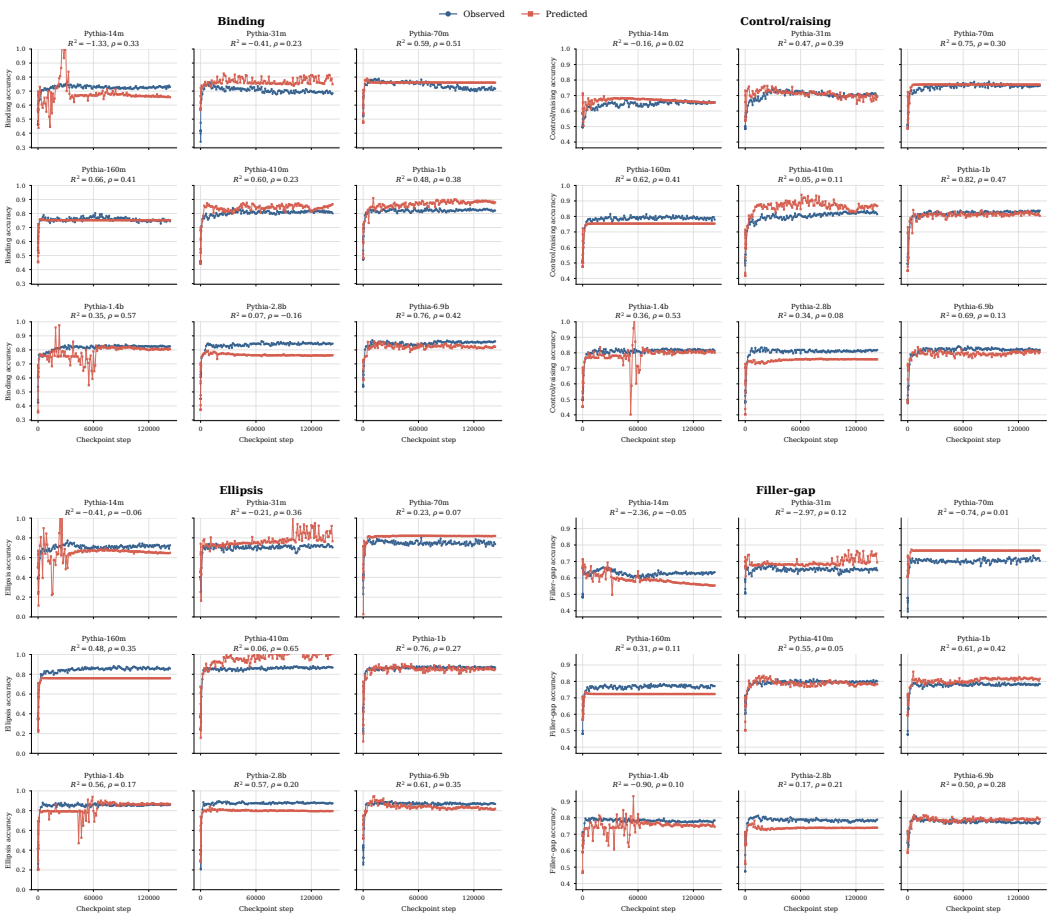


Figure 16: **Syntactic-dependency trajectories.** Observed and predicted trajectories for binding, control/raising, ellipsis, and filler–gap dependencies.

-5.21 to -0.16 at 14m and from -1.84 to 0.31 at 70m. Argument structure at 31m similarly improves from -0.23 to 0.71. The advantage is not universal. Ridge remains better at selected scales, including SVA at 14m, 70m, and 1.4b, and HellaSwag at 410m and 1b. These results suggest that the relationship between representational geometry and behavior is not uniformly linear. Local prediction can be more robust under some held-out-scale shifts, but the preferred mapping depends on both the task and the model population.

Why does kNN perform well here? Under leave-one-scale-out evaluation, the training set contains checkpoints from eight Pythia scales, and the test set contains the checkpoints of the ninth. All scales share the same architecture, training data, and optimization procedure, so a held-out checkpoint usually has close matches among the training checkpoints. Cosine kNN can exploit this by matching each held-out checkpoint to similar training checkpoints and averaging their behavioral scores. For the three tasks in the plot 23, all three neighbors come from a single training scale for 64% of SVA, 81% of argument-structure, and 44% of HellaSwag predictions. This local matching can capture scale-dependent structure that a linear ridge model may miss, which also makes this setting favorable for kNN. Checkpoints from the same training run are highly correlated rather than independent models, and most held-out scales lie between training scales.

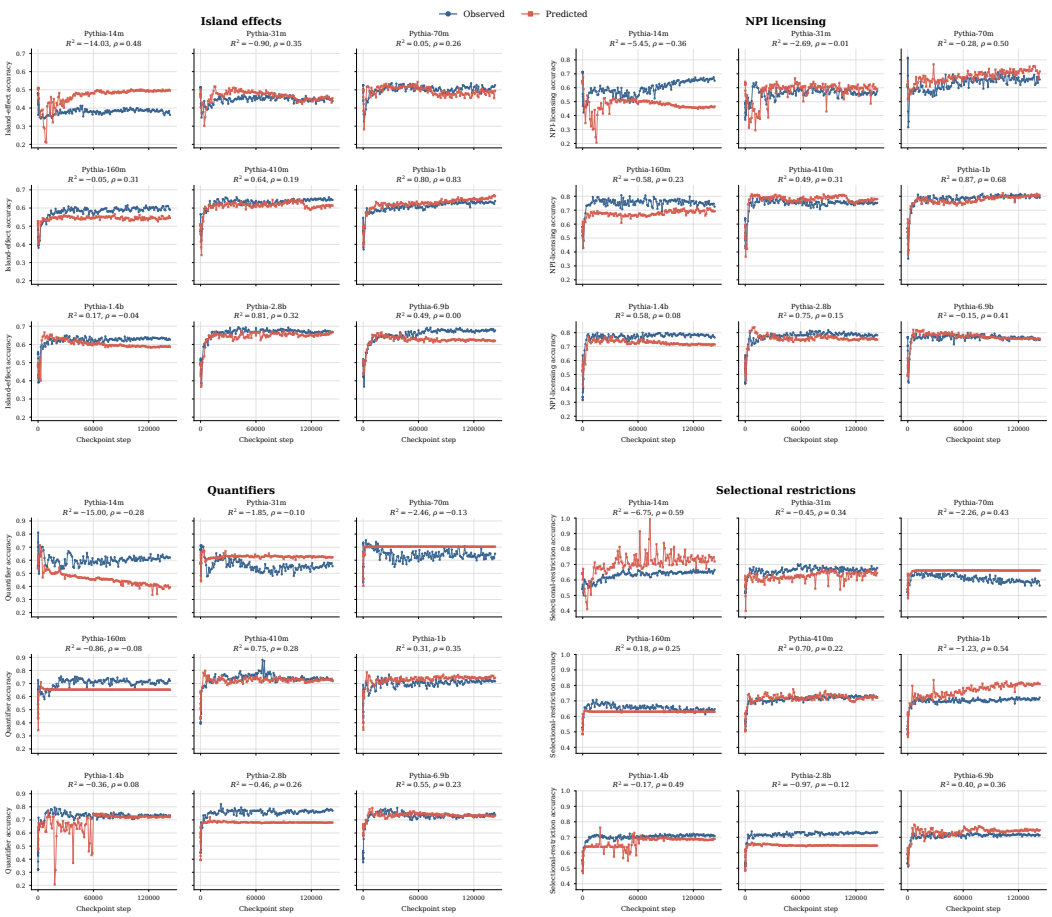


Figure 17: **Licensing and semantic trajectories.** Observed and predicted trajectories for island effects, NPI licensing, quantifiers, and selectional restrictions.

D Brain Experiments

D.1 Experimental Setup

We evaluate 33 mice from the International Brain Laboratory Brain-Wide Map dataset (Angelaki et al., 2025), using one recording session per mouse from the visual contrast-discrimination task. We retain trials from biased blocks and discard contrast levels with fewer than 20 trials. For each of five deterministic trial-split folds, we stratify trials by signed contrast and use 50% of them to construct neural representations and the remaining 50% to estimate the mouse's psychometric slope. The slope is obtained by fitting a four-parameter psychometric function (bias, slope, and two lapse rates) to the held-out choices.

We analyze `probe00` activity from the CA1, DG, LP, PO, and VISa regions. Using only the representation-building trials, we select the 100 neurons whose trial-averaged responses vary most across contrasts. Neural activity is aligned to the behavioral response, binned into ten 40-ms bins spanning the 400 ms after the response, and averaged across trials with the same contrast. Combining nine signed contrast levels with ten time bins produces a 90×100 situation-by-neuron activation matrix for each mouse and fold. We center each neuron across situations, normalize the matrix by its Frobenius norm, and compute pairwise Euclidean distances between situations.

We predict the psychometric slope with ridge regression on standardized RDM features under leave-one-mouse-out evaluation, training on 32 mice and testing on the remaining mouse. The regularization strength is selected from $\{0.1, 1, 10, 100\}$ by leave-one-out cross-validation within the training

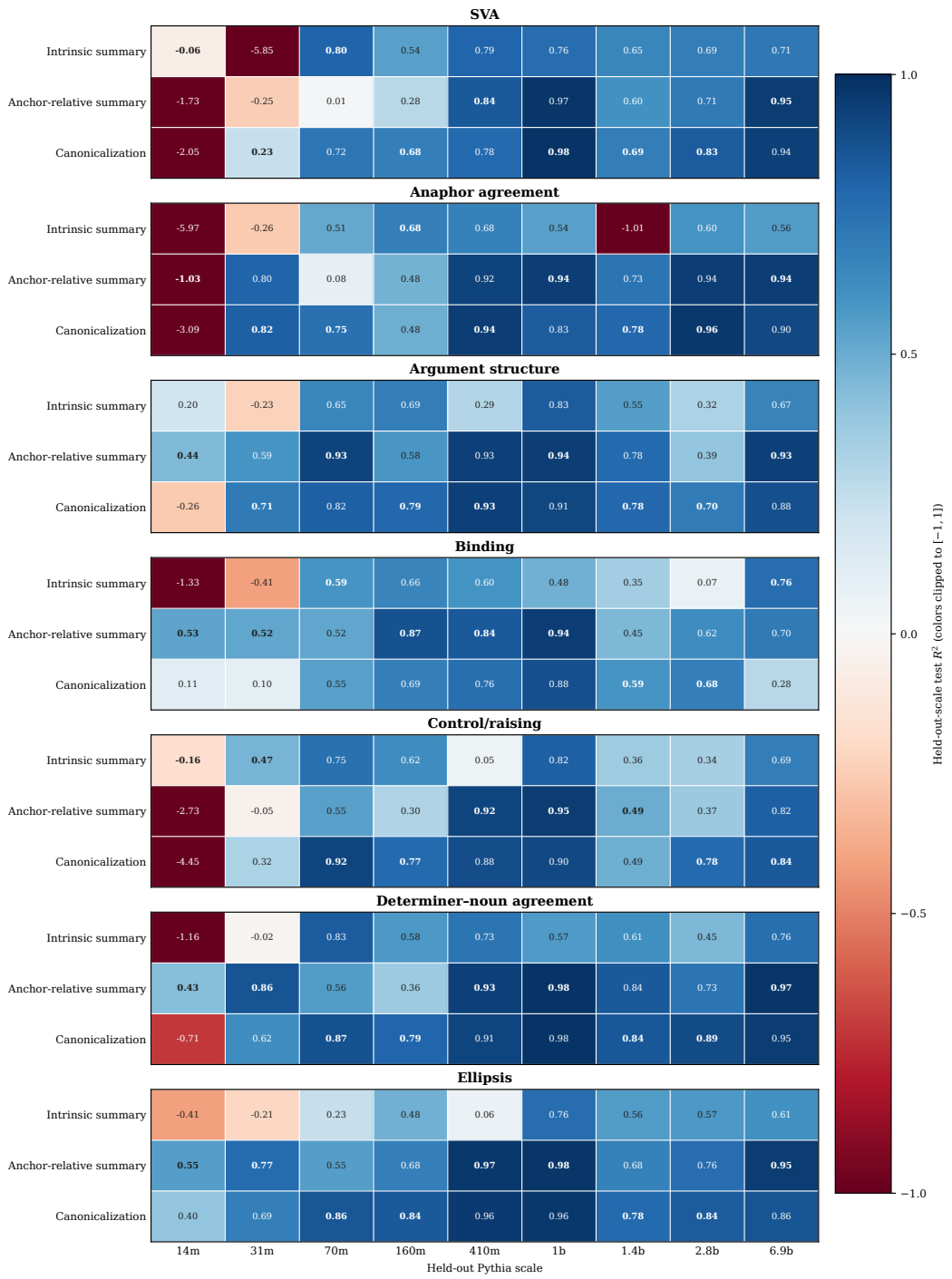


Figure 18: **Comparison of representational descriptions across language tasks (part I).** Held-out-scale test R^2 using the intrinsic summary, anchor-relative summary, or canonicalization. Bold values indicate the best method for each task and scale. Colors are clipped to $[-1, 1]$; annotations show the actual values.

mice, and slopes are z-scored using training-mouse statistics. Our primary metric is pooled z-R^2, which measures variance explained beyond the training-mean baseline; Pearson's r is secondary. Both are computed within each fold and averaged over the five folds.

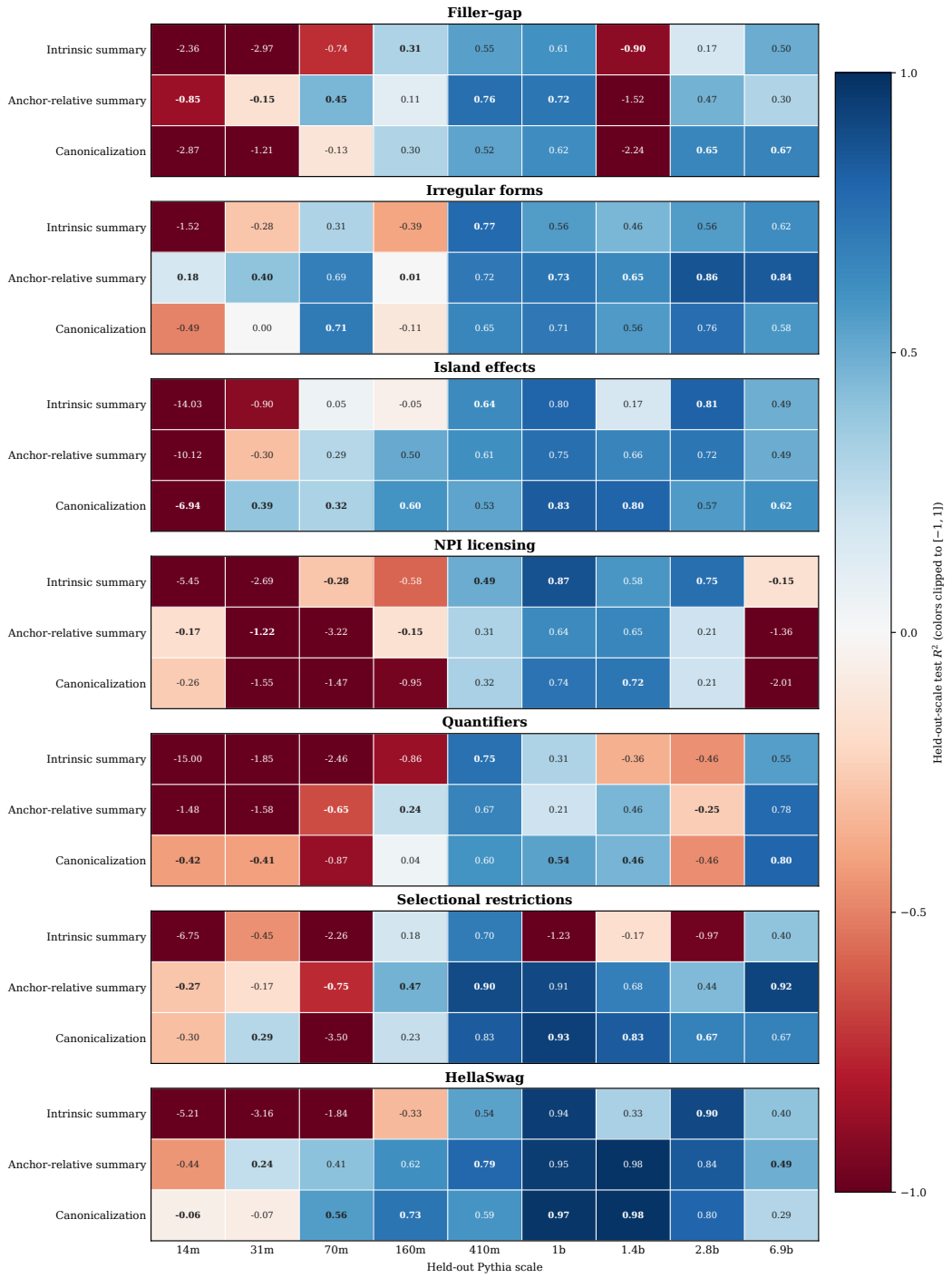


Figure 19: **Comparison of representational descriptions across language tasks (part II).** Continuation of Figure 18 using the same evaluation and visualization conventions.

To study representation scale, we vary temporal resolution across 1, 2, 5, 10, 20, and 50 bins while holding the 400-ms interval and nine contrasts fixed. To study stimulus content, we compare all six balanced five-contrast subsets formed by selecting two nonzero contrast magnitudes from $\{6.25, 12.5, 25, 100\}\%$ together with zero; each subset yields 50 situations.

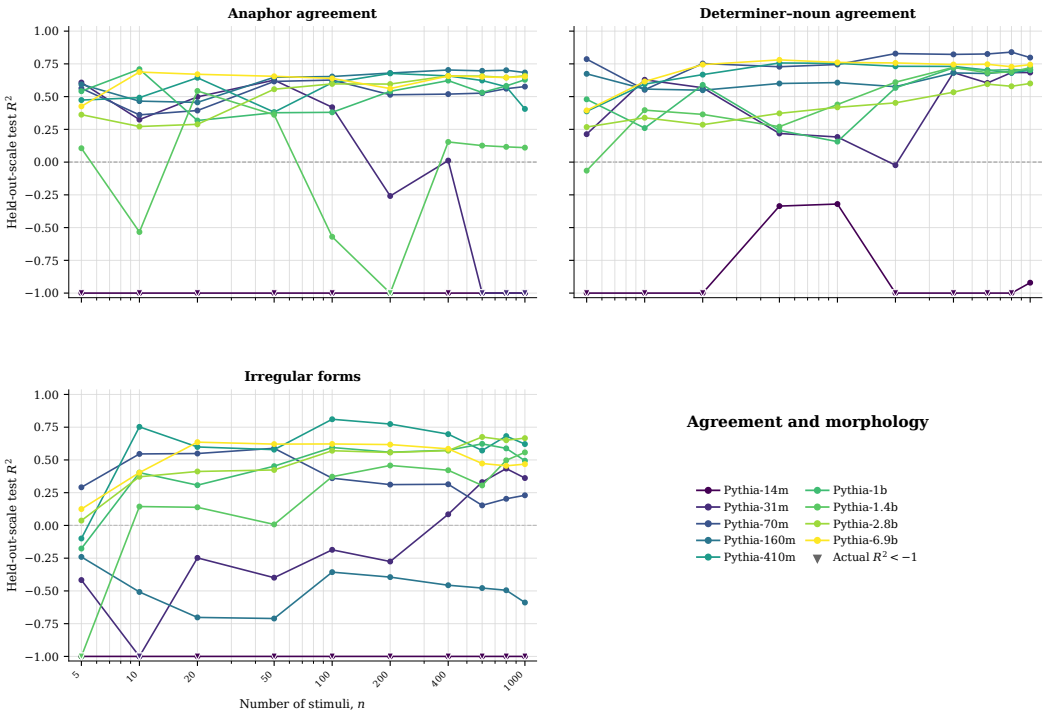


Figure 20: **Agreement and morphology scaling.** Held-out-scale test R^2 for anaphor agreement, determiner–noun agreement, and irregular forms. Downward triangles indicate raw $R^2 < -1$, clipped to -1 for display.

D.2 Additional Brain Results

Temporal resolution. Prediction is at or below chance when each contrast gets only one or two time bins (z-$R^2 < 0$), and it improves as more bins are added. It peaks at the 10-bin setting used in the main text (z-$R^2 = 0.22$, $r = 0.48$) and levels off at 20–50 bins (Fig. 25). This suggests that the behaviorally relevant geometry lies in how responses evolve over time, not in time-averaged activity.

Contrast subsets. All six balanced subsets predict slope about as well as the full nine-contrast set (z-R^2 = 0.17–0.30, r = 0.39–0.53), and their error bars overlap substantially (Fig. 26). No particular contrast range drives the effect: the predictive signal is present whether the representation summary is built from hard, mixed, or easy stimuli.

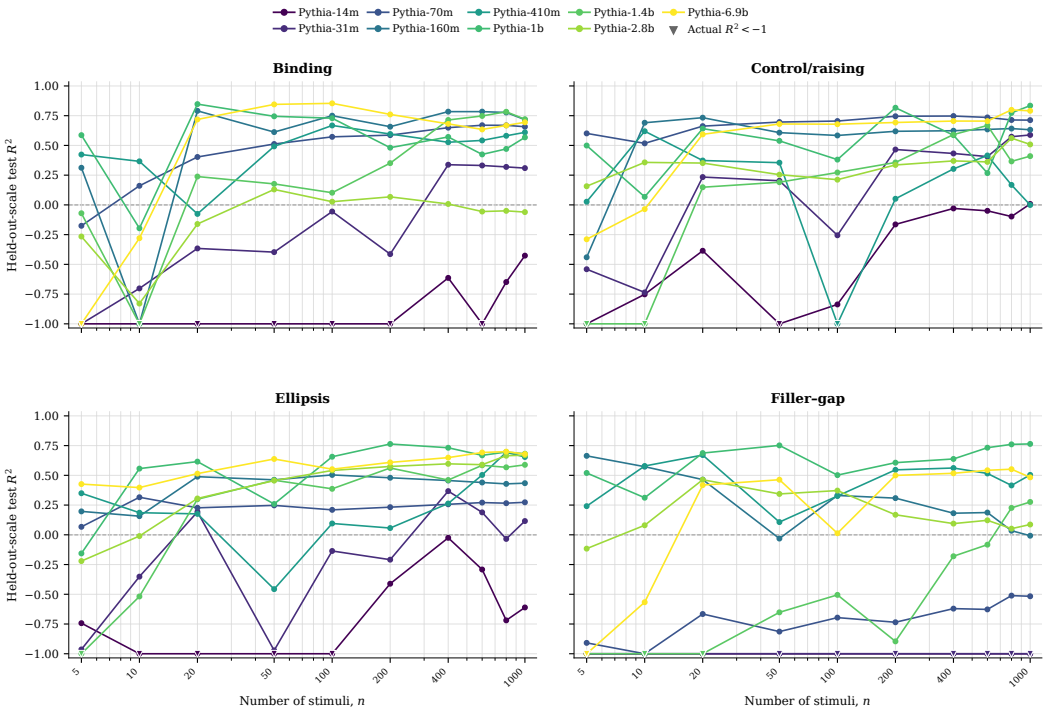


Figure 21: **Syntactic-dependency scaling.** Results for binding, control/raising, ellipsis, and filler–gap dependencies. Plotting conventions follow Figure 20.

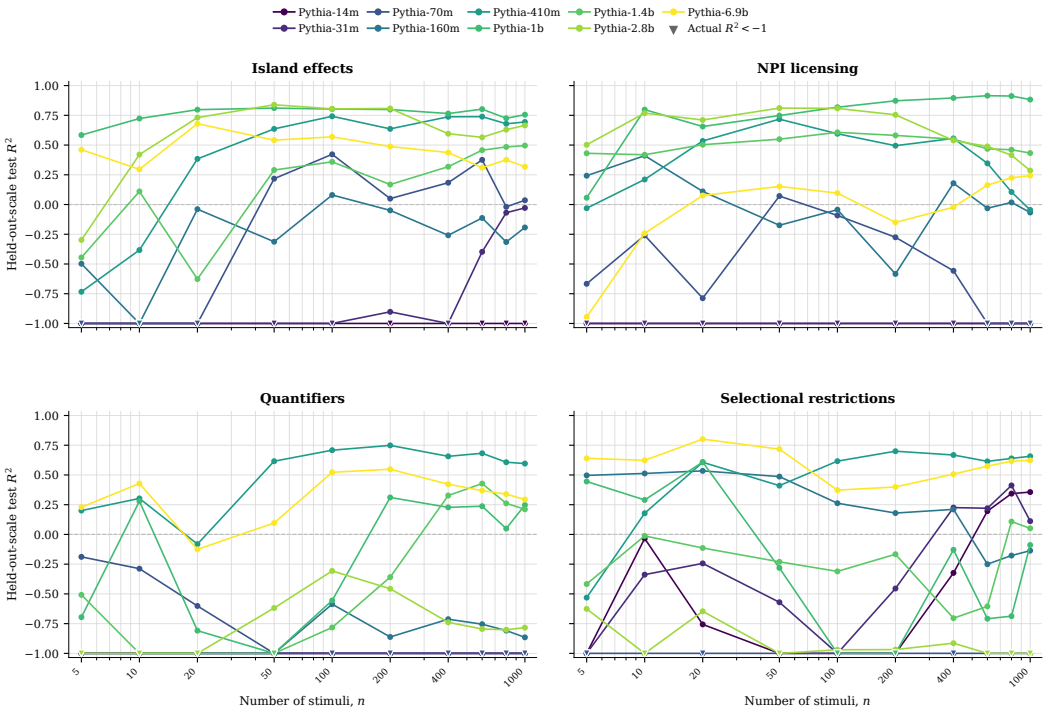


Figure 22: **Licensing and semantic scaling.** Results for island effects, NPI licensing, quantifiers, and selectional restrictions. Plotting conventions follow Figure 20.

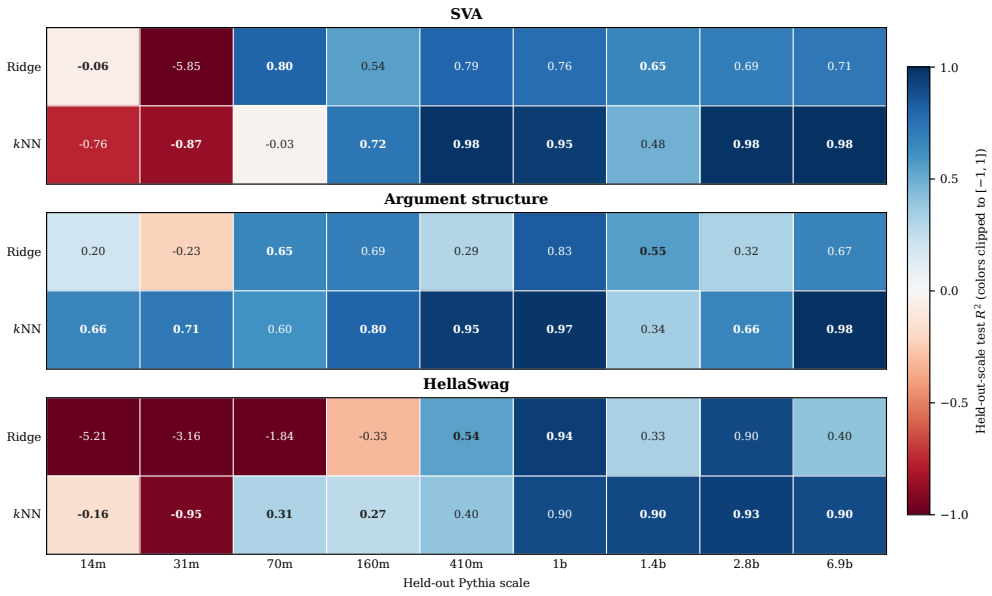


Figure 23: **Ridge versus kNN prediction from language-model RDMs.** Each cell reports held-out-scale test R^2 using matched RDM features constructed from 200 stimuli. Columns indicate the Pythia scale excluded during leave-one-model-size-out evaluation. We compare ridge regression with kNN using $k = 3$, cosine distance, and uniform weighting. Bold values identify the better predictor for each task and held-out scale. Colors are clipped to $[-1, 1]$. Higher is better.

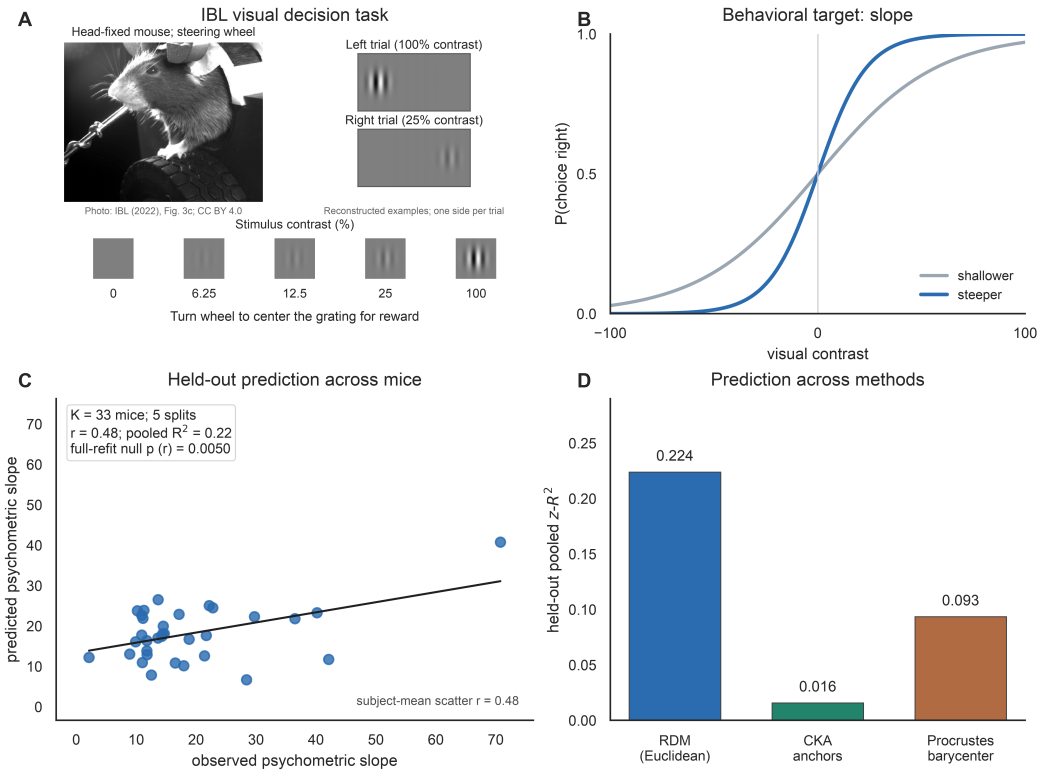


Figure 24: **Predicting individual visual decision sensitivity from neural representations. (A)** In the International Brain Laboratory task, a head-fixed mouse turns a wheel to bring a visual grating, presented on either the left or right, to the screen center for reward. Stimulus contrast varies across trials. Photograph: International Brain Laboratory, Fig. 3c, CC BY 4.0 (`https://doi.org/10.6084/m9.figshare.19694452.v1`). **(B)** Illustrative psychometric curves relate signed stimulus contrast to rightward-choice probability. The fitted slope quantifies behavioral sensitivity to contrast and serves as the prediction target. **(C)** Observed versus predicted psychometric slopes for 33 mice using ridge regression on neural RDM features, evaluated with leave-one-mouse-out cross-validation. Neural representations and behavioral targets are estimated from separate trial halves. Each point averages one mouse's observed and held-out predicted slopes across five repeated trial splits; the black line is a least-squares fit. Split-averaged performance is Pearson's $r = 0.48$ and pooled $z\text{-}R^2 = 0.22$. **(D)** Comparison of intrinsic summaries (Euclidean RDM), anchor-relative summaries (CKA), and canonicalized representations (Procrustes barycenter). Split-averaged pooled $z\text{-}R^2$ values are 0.224, 0.016, and 0.093, respectively.

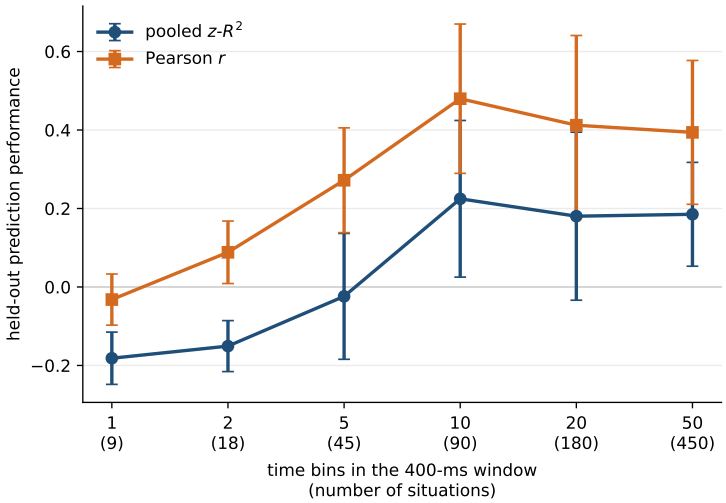


Figure 25: **Temporal-resolution scaling (mice).** Held-out slope prediction (pooled z-R^2 and Pearson r, mean $\pm$ SEM over 5 splits) as the 400-ms response window is divided into more time bins. Numbers in parentheses give the number of situations. The main text uses 10 bins.

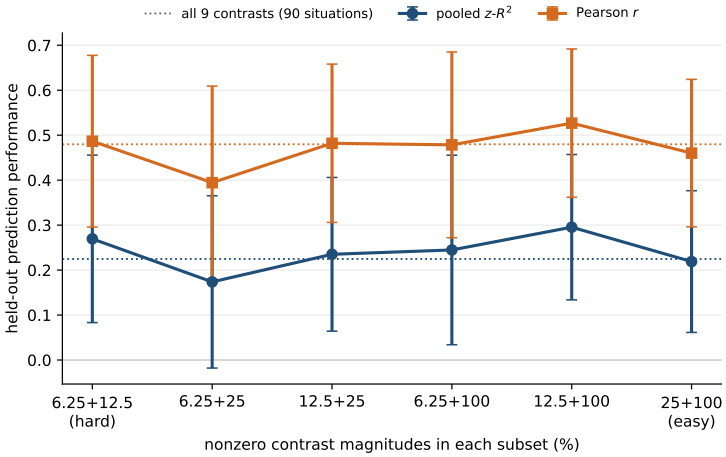


Figure 26: **Contrast subsets (mice).** Held-out slope prediction for six balanced subsets, each containing zero contrast plus two nonzero magnitudes on both sides (50 situations each), ordered from hard to easy. Dotted lines show the all-nine-contrast result from the main text. Mean $\pm$ SEM over 5 splits.