Meta-Learning Captures Human-Like Geometric Sensitivity

Authors Anonymized for Review
Institution(s) Anonymized for Review

Abstract

Humans show systematic biases in geometric perception: shapes with higher *regularity* (symmetry, right angles, parallel sides) and *simplicity* (fewer elements) are easier to recognize. These biases have been difficult to replicate in neural networks, leading to the hypothesis that the human visual system uniquely employs a "geometric language of thought"—discrete symbolic machinery for representing shapes—that cannot be reproduced by neural networks. We show that neural networks trained via meta-learning can reproduce these biases without symbolic primitives or massive pretraining. Training on concepts with fixed complexity yields human-like regularity sensitivity, whereas training on concepts with varying complexity yields simplicity sensitivity. This dissociation suggests that meta-learning induces priors that exploit the most diagnostic dimension of variation in the training environment. More broadly, our results suggest that these perceptual biases can arise from the statistics of the training environment itself, rather than being innately endowed.

Keywords: geometric reasoning; meta-learning; regularity effect; visual cognition

Introduction

Two prominent effects have been documented in the way humans perceive geometric shapes: a *regularity effect*, where shapes with more right angles, parallel sides, and equal lengths are easier to discriminate (Sablé-Meyer et al., 2021); and a *simplicity effect*, where shapes with lower visual complexity are easier to discriminate. Together, these biases have been interpreted as evidence for a uniquely human "language of thought" for geometry, where shapes are mentally represented in terms of discrete primitives that tend to yield shorter descriptions for regular and simple shapes (Sablé-Meyer et al., 2022).

Understanding the origins of these sensitivities to different aspects of geometry is an important and contested scientific challenge. Standard neural networks trained on shape classification do not show human-like error patterns (Sablé-Meyer et al., 2021), leading some to argue that geometric intuition reflects innately symbolic cognitive machinery. Recent work has shown that large pre-trained vision models (e.g., DINOv2 and CLIP) exhibit human-like geometric biases (Campbell et al., 2024), and that neural networks can approximate human performance when trained at very large scale (Pajot et al., 2025). This raises a key question: is massive scale the only route to human-like geometric sensitivity?

We propose that meta-learning—learning to learn across related tasks (Hospedales et al., 2021; Thrun & Pratt, 1998)—provides a mechanism for acquiring human-like geometric biases with far less data. Critically, we find that the specific bias that emerges depends on the structure of the learning domain. In a *constrained* feature space where complexity is controlled and regularity is the primary dimension of variation, meta-trained vision models (convolutional and vision transformer networks) capture human-like regularity sensitivity better than pre-trained and symbolic baselines. In a *diverse* feature space in which both complexity and regularity vary, meta-trained convolutional neural networks instead show a simplicity sensitivity, like humans, whereas meta-trained vision transformers, symbolic models, and pre-trained convolutional neural networks do not.

This dissociation highlights the role of data distributional properties in the emergence of geometric sensitivity: simplicity and regularity effects arise largely as a function of which feature carries the most variance in the training data, holding other factors constant. When the feature space is small and constrained, subtle geometric properties like parallelism and right angles become diagnostic; when the space is large and diverse, overall visual complexity becomes more salient. More generally, our results suggest that meta-learning can induce visual priors that reflect the most diagnostic dimension of variation in the training task distribution—and that this adaptive sensitivity mirrors human geometric perception.

Background

Geometric Biases in Human Cognition

Sablé-Meyer et al. (2021) demonstrated that humans show systematic sensitivity to geometric regularity in an odd-one-out task with quadrilaterals. Participants viewed arrays of nine shapes—eight from one category (e.g., squares) and one slightly distorted exemplar (produced by dragging one vertex outward)—and identified the odd one out. Error rates varied systematically across shape categories, with more regular shapes (those with more right angles, parallel sides, and equal side lengths) yielding lower errors.

This regularity effect was not observed in baboons performing the same task or in pre-trained convolutional neural networks, leading the authors to propose that geometric regularity sensitivity reflects a uniquely human "language of thought" for geometry (Sablé-Meyer et al., 2022). Under this view, shapes are mentally represented as compositions of discrete geometric primitives. Relatedly, humans also show a

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

simplicity effect: shapes with fewer visual elements are easier to discriminate, consistent with minimum description length (MDL; Chater and Vitányi 2003).

Neural Networks for Visual Tasks

We use two kinds of neural networks that have previously been developed for computer vision: convolutional neural networks and vision transformers. Convolutional neural networks (CNNs) are multi-layered neural networks that process pixel-level visual input (LeCun et al., 2015). Inspired by biological vision (Hubel & Wiesel, 1959), early layers learn local filters whose outputs are pooled to produce translation-invariant, multi-scale representations. CNNs achieved breakthrough performance on image classification (Krizhevsky et al., 2012) and remain a standard architecture for computer vision.

Vision Transformers (ViTs; Dosovitskiy et al. 2021) instead partition an image into fixed-size patches, project each patch into an embedding, and process the resulting sequence with a Transformer encoder (Vaswani et al., 2017). Unlike CNNs, ViTs lack built-in spatial inductive biases such as locality and translation equivariance; they instead learn spatial structure from data via self-attention. We compare both architectures to test whether human-like geometric biases depend on architecture-specific inductive biases or can emerge from meta-learning.

Meta-Learning

Machine learning research has explored two distinct approaches to meta-learning. In both cases, the goal is to create a system that is capable of rapidly adapting to new tasks by drawing upon shared experience across previous tasks. The difference is in how that adaptation occurs and hence what is learned from previous tasks.

In *in-context* meta-learning, a single neural network can adapt to new tasks without changing its weights. For example, a recurrent neural network might learn a complex policy that allows it to respond adaptively to new tasks that are presented to it (Duan et al., 2016; Wang et al., 2017). This is also the kind of meta-learning displayed by large language models, in which a single neural network is trained to capture the distributions over words that appear in different documents and hence to be able to model different conditional distributions effectively (Brown et al., 2020).

In *in-weights* meta-learning, adaptation to new tasks depends on updates to the weights of the neural network, typically via a learning algorithm based on gradient descent. In this case, meta-learning adapts the parameters of that learning algorithm, such as the initial weights that it begins from.

Here, we focus on in-weights meta-learning, as our goal is to characterize the inductive biases that enable people to quickly learn new geometric concepts. In particular, we use Model-Agnostic Meta-Learning (MAML; Finn et al. 2017), which identifies a set of initial weights that support rapid learning. Given a distribution of tasks $p(\tau)$, MAML optimizes the initial weights θ^* such that a small number of gradient steps from that initialization yields good performance across tasks. Formally, MAML minimizes the expected post-adaptation loss:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\tau \sim p(\tau)} \left[\mathcal{L}(\theta - \alpha \nabla_\theta \mathcal{L}(\theta; D_\tau^s); D_\tau^q) \right] \quad (1)$$

where D_τ^s and D_τ^q are the support and query sets for task τ, and α is the inner learning rate. Grant et al. (2018) showed that this objective can be interpreted as learning an implicit prior over model parameters: predictions at zero or a few adaptation steps are prior-dominated, whereas further adaptation becomes increasingly likelihood-dominated. Previous work has shown that MAML can be used effectively to create systems that are capable of rapid learning in domains that are relevant to cognitive science: McCoy and Griffiths (2025) used this approach to distill Bayesian linguistic priors into recurrent neural networks, and Gupta et al. (2025) showed that it enables CNNs to learn the same-different relation where standard training fails.

Methods

To test whether meta-learning can endow neural networks with human-like geometric sensitivity, we train convolutional neural networks and vision transformers (Dosovitskiy et al., 2021) on two sets of geometric concepts from prior work and compare the resulting error patterns to human behavior.

Stimuli and Human Data

We use two stimulus sets. First, we replicate the 10-concept quadrilateral family from Sablé-Meyer et al. (2021). These stimuli vary systematically in geometric regularity, from highly regular (square: 4 right angles, 2 pairs of parallel sides, 2 pairs of equal sides) to highly irregular (rusted hinge: no right angles, no parallel sides, no equal sides), by modulating the internal angles of a base quadrilateral. Second, using the same renderer, we replicate geometric constructions from the Geoclidean framework (Hsu et al., 2022): 37 concepts ranging from basic elements (triangle, angle, parallel lines, perpendicular bisector) to composite configurations (e.g., "ccc" = 3 mutually tangent circles; "llll" = 4 intersecting lines).

Figure 1 shows example stimuli from both concept sets. The category membership task (a) requires learning a concept–label mapping for each concept (shown for the diverse setting with all 47 concepts described below). The odd-one-out task (b) requires identifying the outlier that violates one of the constraints satisfied by the five positive examples.

For each of the 47 categories, we generated training and testing images (128×128 pixels) using Cairo rendering under a single renderer. For quadrilaterals, we created odd-one-out examples via vertex-dragging augmentation; for Geoclidean concepts, we used "near" and "far" variants that break one or more geometric constraints following Hsu et al. (2022). Shapes were rendered as black outlines on white backgrounds. Given that the generator is stochastic, no two examples are exactly the same, and variation is constrained with pre-set degrees of rotation, pixelation, scale, and color variation.

We used two sources of human behavioral data from prior work. For the quadrilateral concepts, mean error rates were taken from French adults (n = 612 trials per shape) in an odd-one-out task reported by Sablé-Meyer et al. (2021). For the Geoclidean concepts, error rates were taken from an online odd-one-out experiment (n = 286 adult participants) conducted by Budny et al. (2025).

Models and Training

We compare two architectures under two training regimes: standard pre-training with Adam (Kingma & Ba, 2014) and meta-learning with first-order MAML ("meta-training"). The full algorithm for meta-training with MAML appears in the Appendix.

CNN: A 4-layer convolutional network (Conv4) with 64 filters per layer, batch normalization, ReLU activations, and 2×2 max pooling after each block (∼113K parameters). Following Gupta et al. (2025), who showed that relatively shallow CNNs can meta-learn geometric relations, we evaluated CNNs with 2, 4, and 6 layers and found the 4-layer variant to perform best; we therefore use this architecture throughout.

Vision Transformer (ViT): A ViT-Tiny model with 12 layers, 3 attention heads, hidden dimension 192, and patch size 16 (∼5.5M parameters). We use the HuggingFace implementation without pre-trained weights.

Symbolic Model: As a baseline, we compare against the symbolic geometric-features model of Sablé-Meyer et al. (2021). This model represents each quadrilateral as a 22-dimensional binary feature vector encoding discrete geometric properties: 4 right-angle indicators (whether each internal angle is 90°), 6 parallel-pair comparisons, 6 equal-length comparisons, and 6 equal-angle comparisons. Pairwise shape dissimilarity is computed as the Hamming distance between feature vectors. This model operationalizes the language-of-thought hypothesis—that humans represent shapes as symbolic compositions of geometric primitives—and achieves $\rho = 0.72$ Spearman rank-order correlation with human error patterns on quadrilateral odd one out classification (Table 1).

We train on two setups: quadrilaterals-only (10 concepts) and combined quadrilaterals and geoclidean (47 concepts). In each episode, the model learns binary category membership over a pair of concepts. At test time, we evaluate on odd-one-out with 5 positive exemplars and 1 negative, matching the human experimental setup. Training ran for 100 epochs ($\beta = 0.001$, $\alpha = 0.01$), and we evaluate performance across adaptation steps (K = 0–20).

Metrics

We quantify two geometric properties central to our analysis.

Regularity, following Sablé-Meyer et al. (2022), is the normalized count of discrete geometric constraints present in a shape:

$$\text{Regularity} = \frac{n_{\text{right angles}} + n_{\text{parallel pairs}} + n_{\text{equal pairs}}}{8} \quad (2)$$

where each n counts the corresponding geometric property and 8 is the maximum possible value (attained by a square). Scores range from 0 (rusted hinge) to 1 (square). For Geoclidean concepts, we extend this to a weighted combination of straightness (proportion of lines vs. circles), right angles, and parallel pairs.

Simplicity is operationalized as the inverse of visual complexity, measured as the number of visible strokes in the rendered image. Concepts range from 2 strokes (e.g., two lines forming an angle) to 5 strokes (e.g., three triangles, a circle, and a line).

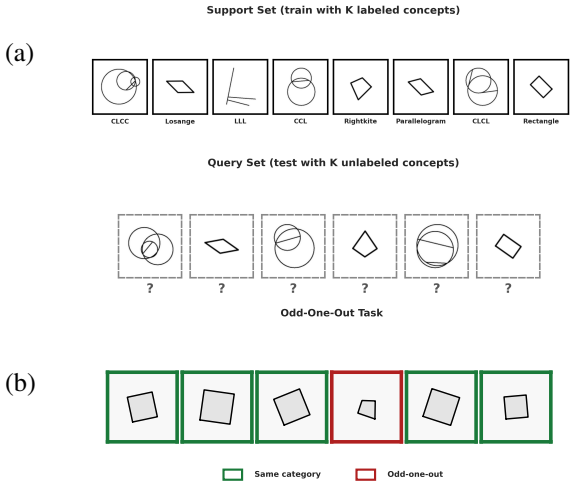

Figure 1: Meta-learning for geometric concepts. (a) Training is based on a 47-way classification problem, where the model learns concept–label mappings by adapting on support sets and evaluating on query sets (one per episode). (b) Testing on a 10-way odd-one-out task with quadrilaterals, where the model selects the image that violates the concept (matching the paradigm from Sablé-Meyer et al., 2021).

These metrics capture distinct aspects of geometric structure: regularity reflects internal geometric constraints, while simplicity reflects overall visual complexity. In the quadrilateral subset, all shapes have four sides, so complexity is controlled and regularity varies systematically. Across the full 47 concept dataset, both dimensions vary. We computed Spearman rank-order correlations between these metrics and human and model accuracies, finding the predicted positive correlation between simplicity, regularity, and accuracy.

Results

We evaluate meta-learning in two settings that differ in featural complexity: a *constrained* setting where complexity is controlled and a *diverse* setting where complexity varies. In both settings, standard pre-training performs at (or near) chance, whereas meta-learning enables robust geometric concept learning.

Overall Error Correlations: Regularity in Quadrilaterals. We meta-train models on 10 quadrilateral concepts

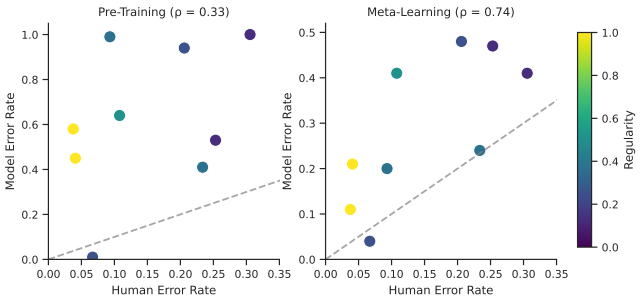

Figure 2: **Overall Error Correlations: Model Vs. Humans** Scatter plots comparing human and model overall error rates for a pre-trained CNN (left) and a meta-trained CNN (right) evaluated on the quadrilaterals task. Each dot denotes an average of the judgements of the 286 human participants and the average judgements of 10 cross-validated models on outlier detection of each of the 10 quadrilaterals. Points are colored by geometric regularity (yellow = high, purple = low). The meta-trained CNN captures the human regularity effect; the pre-trained CNN does not. The dashed line indicates perfect alignment.

Table 1: Human–model alignment in the constrained setting (quadrilaterals only). Models were trained on 10 quadrilateral concepts. The symbolic model from Sablé-Meyer et al. (2022) uses 22 hand-crafted geometric features; Meta-CNN is evaluated at different adaptation steps. $^{*}p < 0.05$, $^{**}p < 0.01$.

Model	**Adapt.**	n	ρ	p
Symbolic (Sablé-Meyer)	—	10	0.72	0.02^{*}
Meta-CNN	0 steps	10	0.74	0.01^{*}
	2 steps	10	0.87	0.001^{**}
	5 steps	10	0.79	0.006^{**}
	10 steps	10	0.57	0.08

for which regularity is the primary dimension of variation. Figure 2 shows that the meta-trained convolutional neural network (Meta-CNN) aligns strongly with human error patterns at early adaptation, whereas the pre-trained CNN does not. Meta-CNN also *outperforms the hand-crafted symbolic model* of Sablé-Meyer et al. (2022) (Table 1). Alignment decreases with further adaptation, consistent with the idea that early steps are prior-dominated.

Both architectures show regularity sensitivity in this setting (Figure 3, top row), suggesting that this effect does not depend strongly on architecture. Moreover, Meta-CNN shows substantially higher alignment with human error patterns than the pre-trained CNN (Figure 2; Table 1).

Diverse Setting: Simplicity Across 47 Concepts. We next train on all 47 concepts, where both regularity and complexity vary. Table 2 reports human–model correlations by subset. Meta-CNN aligns significantly with human errors on Geoclidean concepts, whereas the meta-trained ViT shows no human alignment despite higher accuracy.

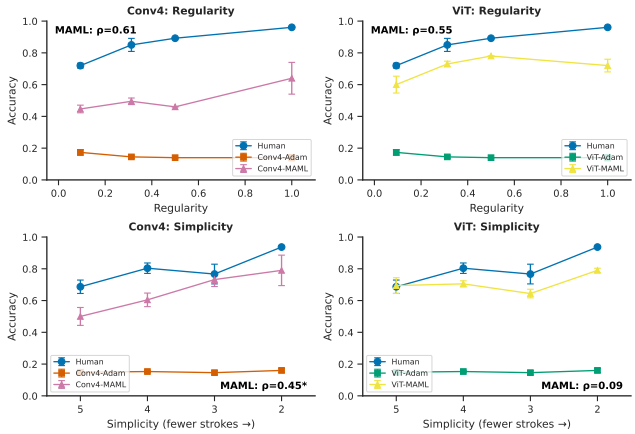

Figure 3: **Regularity and simplicity sensitivity: Human vs. Pretrained vs. Metatrained.** Top: Regularity vs. accuracy on quadrilaterals (constrained setting). Both the meta-trained CNN ($\rho = 0.61$) and ViT ($\rho = 0.55$) show positive correlations with regularity. Bottom: Simplicity vs. accuracy on 47 concepts (diverse setting). The meta-trained CNN shows a significant simplicity effect ($\rho = 0.45^{*}$), whereas the meta-trained ViT shows no effect ($\rho = 0.09$) despite higher accuracy. Each reported result is the average across 10 cross-validated model seeds. $^{*}p < 0.05$.

To test whether simplicity rather than regularity drives this alignment, we computed correlations between each geometric dimension and model accuracy (Table 3). Meta-CNN shows a significant *simplicity* effect ($\rho = 0.45$, $p < 0.01$) but no regularity effect, whereas Meta-ViT shows neither. This suggests that when complexity varies, it becomes the diagnostic dimension—but only CNNs exploit it in a human-like way.

Discussion

Humans show a sensitivity to geometric concepts that has previously only been captured by symbolic models or large-scale neural networks with extensive training. We investigated whether smaller, simpler neural networks can reproduce the same behaviors when trained via meta-learning, which seeks a set of initial weights that abstract the common elements across a set of tasks. Our results show that meta-learning is extremely effective in inducing appropriate inductive biases in simple neural networks. In the remainder of the paper we consider some of the implications of these results for understanding geometric inductive biases in humans and machines.

Meta-Learning Surpasses Symbolic Accounts

Our central finding is that meta-learning produces human-like geometric biases that *exceed* hand-crafted symbolic models. The symbolic model of Sablé-Meyer et al. (2022) was explicitly designed with 22 geometric features to capture human regularity sensitivity, yet Meta-CNN at early adaptation achieves stronger alignment ($\rho = 0.87$ vs. $\rho = 0.72$; Table 1) without

Table 2: Human–model alignment in the diverse setting (47 concepts). Models were trained jointly on all concepts; correlations were computed between human and model test-time errors by subset and overall. $^{*}p < 0.05$.

Model (Acc.)	Eval. Subset	n	ρ	p
Meta-CNN (63.7%)	Quadrilaterals	10	0.52	0.12
	Geoclidean	37	0.37	0.03*
	All concepts	47	0.31	0.03*
Meta-ViT (69.7%)	Quadrilaterals	10	0.56	0.09
	Geoclidean	37	−0.07	0.70
	All concepts	47	−0.01	0.95

Table 3: Fine-grained analysis: Which geometric dimension predicts model accuracy in the diverse setting? Meta-CNN shows significant simplicity sensitivity but no regularity sensitivity, whereas Meta-ViT shows neither. $^{**}p < 0.01$.

Model	Dimension	ρ	p
Meta-CNN	Regularity	+0.06	0.69
	Simplicity	+0.45	0.001**
Meta-ViT	Regularity	+0.04	0.79
	Simplicity	+0.09	0.54

explicit geometric feature engineering. Alignment peaks at early adaptation steps and declines with further adaptation, consistent with the interpretation that MAML's initialization acts as a learned prior (Grant et al., 2018) and that human-like behavior reflects this prior before it is overwritten by task-specific learning.

The symbolic model has no analogue for this dynamic: it assigns a fixed dissimilarity regardless of exposure. This result parallels findings in other cognitive domains: McCoy and Griffiths (2025) showed that MAML-trained recurrent networks match Bayesian learners in acquiring formal languages, and Gupta et al. (2025) found that meta-learning enables CNNs to learn the same-different relation where standard training fails. Across language, relational reasoning, and geometric perception, meta-learning appears to be a general mechanism for instilling human-like inductive biases into neural networks.

Adaptive Sensitivity to Task Structure

A central question is why meta-learning yields regularity sensitivity in one setting but simplicity sensitivity in another. It is worth emphasizing the episodic query/support split of our data that structures the learning experience episodically rather than as a randomly interleaved batch like in pre-training. By grouping similar tasks repeatedly in the support and query sets of each episode, we provide extra signal to the model via this re-ordering of the data alone.

In doing so, we argue that MAML emphasizes whichever geometric dimension is most diagnostic in the task distribution. In the constrained setting, all shapes have four sides, so visual complexity is controlled and regularity (right angles, parallelism, equal lengths) becomes the primary axis along which concepts differ. Meta-learning exploits this structure, yielding strong regularity–accuracy correlations for both architectures (Figure 3, top). In the diverse setting, concepts range from simple angles to complex multi-element configurations, and complexity becomes the dominant source of variation. Here, Meta-CNN shifts toward simplicity sensitivity ($\rho = 0.45$, $p < 0.01$; Table 3) while showing no regularity sensitivity ($\rho = 0.06$).

This transition reflects a broader principle: learners—human or artificial—become sensitive to whichever feature is most informative in their environment. When complexity is held constant, regularity differentiates concepts; when complexity varies, it can overwhelm subtler geometric properties. By optimizing over a distribution of tasks, meta-learning encodes this environmental structure into the model initialization (see the Appendix for a formal Bayesian account).

The Role of Architectural Inductive Bias

The CNN–ViT dissociation suggests that human-like biases require not only structured learning but also compatible architectural inductive biases. In the constrained setting, both architectures show regularity sensitivity, indicating that MAML can induce regularity priors even when inductive biases differ. In the diverse setting, however, only Meta-CNN captures the simplicity effect; Meta-ViT achieves higher overall accuracy (69.7% vs. 63.7%) but shows no correlation with human errors (Table 2).

One explanation is that CNNs have built-in preferences for low-complexity visual features due to locality and weight sharing (Goldblum et al., 2024)—properties that MAML can amplify into a simplicity bias. ViTs, which rely on global self-attention, lack these spatial priors and must learn spatial structure from data. The result may be effective but "non-human" representations: ViTs solve the task without developing the complexity sensitivity that characterizes human geometric perception. This parallels the human visual cortex, which exhibits strong locality constraints (retinotopic organization and local receptive fields) more akin to the inductive biases of CNNs than global self-attention.

Many Paths to Geometric Sensitivity

A key motivation for this work is the claim that geometric regularity sensitivity is uniquely human. Sablé-Meyer et al. (2021) found that neither baboons nor standard neural networks exhibit the regularity effect, and argued that it reflects an innate "language of thought" for geometry—discrete symbolic machinery that is absent in other species and irreproducible by neural networks (Sablé-Meyer et al., 2022). Our results challenge this conclusion directly: meta-learned CNNs not only reproduce the regularity effect but also provided a closer fit to human behavior than the symbolic model designed to formalize the language-of-thought account. This suggests that geometric sensitivity is not a signature of uniquely human

symbolic cognition, but a learnable property that can emerge from the structure of the training environment an adaptable, intelligent agent is placed within.

These results support the idea that there are multiple paths to human-like geometric biases. Symbolic models achieve them through hand-crafted geometric features (Sablé-Meyer et al., 2022). Large-scale pre-trained vision models (e.g., DINOv2 and CLIP) achieve them through exposure to internet-scale visual data (Campbell et al., 2024). Meta-learning achieves them through structured training on a small task distribution, using ∼1000× less data than the pre-trained models and no feature engineering. The convergence of these approaches on similar behavioral signatures suggests that geometric sensitivity is not the result of a single cognitive mechanism, but a functional outcome that can be realized by multiple computational strategies—provided they supply appropriate inductive biases, whether through architecture, data scale, or learning structure.

Limitations and Future Directions

Several limitations warrant discussion. First, we make the explicit choice to train small models from scratch on highly simplified geometric shapes for stimulus and training control over the emergence of the regularity and simplicity biases. However, humans of course experience much richer, naturalistic visual stimuli and we did not test these networks on naturalistic images in this work. Secondly and relatedly, our simplicity metric (stroke count) is coarse; future work could test more principled complexity measures, such as entropy-based measures. Second, while we show that meta-learning *can* produce human-like biases, we do not claim it is the mechanism humans use—only that it provides a sufficient computational account. Comparing meta-learning trajectories to human developmental data could help evaluate this stronger claim. Finally, probing the internal representations of meta-learned models—for example, whether they discover explicit parallel-line or right-angle detectors—could clarify whether learned features resemble the geometric primitives posited by the language-of-thought hypothesis.

Conclusion

Our results show that structured learning across related tasks, combined with appropriate architectural constraints, is sufficient to produce human-like geometric sensitivity. The regularity-to-simplicity transition across our two settings points to a shared computational principle: learners exploit the most diagnostic dimension of variation in the task distribution. This reframes geometric intuition not as uniquely human symbolic machinery, but as an adaptive response to the structure of perceptual experience.

Appendix

MAML Algorithm

Algorithm 1 MAML for Geometric Concept Learning

Require: Concept set $C = \{c_1, \ldots, c_K\}$, inner learning rate α, outer learning rate β

Initialize parameters θ randomly
for each meta-training iteration **do**
  Sample a task batch: $\tau_i = (c_i^+, c_i^-)$ pairs from C
  for each task τ_i **do**
    Sample support D_i^s and query D_i^q from c_i^+, c_i^-
    Adapt: $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}(\theta; D_i^s)$
    Compute query loss: $\mathcal{L}_i = \mathcal{L}(\phi_i; D_i^q)$
  end for
  Meta-update: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i$
end for
return Learned initialization θ^*

Bayesian Derivation of the Simplicity Prior

Here we formalize how meta-learning induces a simplicity prior through implicit Bayesian marginalization, following Grant et al. (2018).

Setup Consider classifying an image x into one of K geometric concepts $C = \{c_1, \ldots, c_K\}$. Each concept c defines a family of shapes through geometric constraints, parameterized by $\theta_c \in \Theta_c \subseteq \mathbb{R}^{d_c}$, where d_c is the number of free parameters remaining after the constraints are applied. More regular concepts impose more constraints and therefore have fewer free parameters:

Concept	Constraints	d_c
Square	4 right angles, 4 equal sides	3
Rectangle	4 right angles, 2 pairs equal	4
Parallelogram	2 pairs parallel, 2 pairs equal	5
Random quadrilateral	None	8

Marginal Likelihood Favors Simplicity The Bayesian posterior over concepts is:

$$p(c \mid x) = \frac{p(x \mid c)\, p(c)}{\sum_{c'} p(x \mid c')\, p(c')} \tag{3}$$

where the likelihood $p(x \mid c)$ is obtained by marginalizing over the concept's parameter space:

$$p(x \mid c) = \int_{\Theta_c} p(x \mid \theta_c)\, p(\theta_c \mid c)\, d\theta_c \tag{4}$$

This integral reflects Bayesian Occam's razor (MacKay, 2003): concepts with fewer parameters (lower d_c) concentrate probability mass in a smaller region of image space, yielding a higher marginal likelihood for their typical instances. Intuitively, a square "wastes" less probability mass on images it cannot generate than a random quadrilateral does.

Under a uniform prior $p(\theta_c \mid c)$ over a parameter space of volume V_c, the marginal likelihood scales as:

$$p(x \mid c) \propto V_c^{-1} \propto L^{-d_c} \tag{5}$$

where L is a characteristic length scale. Thus $\log p(x \mid c) \approx -d_c \log L + \text{const}$, and simpler concepts are exponentially favored.

Adaptation Steps as Prior-Likelihood Tradeoff At test time, adapted parameters after K gradient steps can be written as:

$$\phi^{(K)} = \theta^* - \alpha \sum_{k=0}^{K-1} \nabla_\phi \mathcal{L}(\phi^{(k)}; D^s) \tag{6}$$

This has a natural interpretation:

- $K = 0$: Predictions use θ^* directly, reflecting only the learned prior.
- K small (1–3): The model balances the prior with task-specific evidence, approximating the Bayesian posterior.
- K large (> 5): Gradient descent moves parameters far from θ^* and may overfit to task-specific features; the simplicity prior is "forgotten."

This explains the empirical finding that human-model correlation peaks at $K = 2$ and decreases with further adaptation (Table 1): early adaptation is prior-dominated (showing human-like simplicity/regularity effects), while extended adaptation becomes likelihood-dominated (task-specific).

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

Budny, N., Ghods, K., Campbell, D., Marjieh, R., Joshi, A., Kumar, S., Cohen, J. D., Webb, T. W., & Griffiths, T. L. (2025). Visual serial processing deficits explain divergences in human and VLM reasoning. *arXiv preprint arXiv:2509.25142*.

Campbell, D., Kumar, S., Giallanza, T., Griffiths, T. L., & Cohen, J. D. (2024). Human-like geometric abstraction in large pre-trained neural networks. *arXiv preprint arXiv:2402.04203*.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 1126–1135.

Goldblum, M., Finzi, M., Rowan, K., & Wilson, A. G. (2024). The no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. *Proceedings of the 41st International Conference on Machine Learning*.

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. *International Conference on Learning Representations*.

Gupta, M., Rane, S., McCoy, T., & Griffiths, T. L. (2025). Convolutional neural networks can (meta-)learn the same-different relation. *arXiv preprint arXiv:2503.23212*.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169.

Hsu, J., Wu, J., & Goodman, N. D. (2022). Geoclidean: Few-shot generalization in Euclidean geometry. *Advances in Neural Information Processing Systems*, *35*, 39007–39019.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

McCoy, R. T., & Griffiths, T. L. (2025). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, *16*, 5765.

Pajot, M., Morfoisse, T., Sablé-Meyer, M., Lakretz, Y., & Dehaene, S. (2025). Can neural networks model the human perception of geometric shapes? *OSF Preprints*.

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527.

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences*, *118*(16), e2023123118.

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2017). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.