# What Makes a Good Example? Modeling Exemplar Selection with Neural Network Representations

**Fanxiao Wani Qiu***
Department of Psychology
University of Southern California
fanxiaoq@usc.edu

**Oscar Leong***
Department of Statistics and Data Science
University of California, Los Angeles
oleong@stat.ucla.edu

**Alexander LaTourrette**
Department of Psychology
University of Southern California
latourre@usc.edu

## Abstract

Teaching requires distilling a rich category distribution into a small set of informative exemplars. Although prior work shows that humans consider both representativeness and diversity when teaching, the computational principles underlying these tradeoffs remain unclear. We address this gap by modeling human exemplar selection using neural network feature representations and principled subset selection criteria. Novel visual categories were embedded along a one-dimensional morph continuum using pretrained vision models, and selection strategies varied in their emphasis on prototypicality, joint representativeness, and diversity. Adult participants selected one to three exemplars to teach a learner. Model–human comparisons revealed that strategies based on joint representativeness, or its combination with diversity, best captured human judgments, whereas purely prototypical or diversity-based strategies performed worse. These results highlight the potential utility of dataset distillation methods in machine learning as computational models for teaching.

**Keywords:** computational modeling; dataset distillation; teaching; exemplar selection; feature representation

Teaching is a process central to cumulative culture. Infants are sensitive to teaching, treating information taught by adults as generalizable and culturally relevant (Csibra & Gergely, 2011). By preschool age, children themselves can engage in selective teaching, prioritizing the transmission of optimally helpful information (Qiu et al., 2025). This sensitivity to teaching, both as benefactors and beneficiaries, supports the accumulation of knowledge across generations and underlies humans' uniquely rich representational systems, including language, art, and science. An important part of teaching is the selection of informative examples (Shafto et al., 2014). For instance, when teaching a child about birds, one might weigh both prototypicality (e.g., robins and sparrows are typical examples) and diversity (e.g., penguins and ostriches illustrate variation within the category) when deciding which exemplars best convey the category structure. A substantial body of literature in category-based induction has examined how learners evaluate examples (Heit, 2000). People draw stronger inferences from typical category members than atypical ones (e.g., robins vs. ostriches for birds) (Osherson et al., 1990), from diverse sets as opposed to similar sets (robins and ostriches vs. robins and sparrows) (Heit, 2000; Rhodes et al., 2008), and from larger samples than smaller ones (Nisbett et al., 1983).

Despite the importance of selecting good examples for successful teaching, there has been relatively little systematic work that examines *how* humans select examples. Prior work suggests that adults do not simply select at random. Shafto and colleagues formalized teaching as a process of Bayesian inference, in which teachers choose examples to maximize a learner's posterior belief in the accurate hypothesis, rather than to reflect the underlying data distribution (Shafto et al., 2014). Related work shows that adults prefer informative positive examples when teaching concepts (Avrahami et al., 1997), and that both children and adults favor diverse over similar examples when teaching others about a novel property in animals (Rhodes et al., 2010). These findings suggest that human teachers are sensitive to the informativeness of examples. However, the computational principles governing how humans resolve tradeoffs between typicality and diversity in selecting examples remain poorly understood.

A parallel set of questions has been explored in machine learning under dataset distillation. Motivated by computational considerations and the observation that large datasets are often redundant, this body of work asks whether a small subset of curated examples can capture the essential structure of a larger dataset. Early works include core-set construction (Har-Peled & Kushal, 2007; Tsang et al., 2005), facility location techniques (Mirchandani & Francis, 1990), and instance selection methods (Olvera-López et al., 2010), which seek to select a subset of training examples such that models trained on this subset achieve performance comparable to those trained on the full dataset. These techniques continue to be explored in the context of active learning (Li et al., 2023; Sener & Savarese, 2018; Wei et al., 2015). Building on these ideas, recent work has explored the synthetic generation of small training sets that are explicitly optimized to encode the information contained in a larger dataset (Wang et al., 2018). This line of work, commonly known as dataset distillation (Lei & Tao, 2023), has been shown to be beneficial in reducing memorization (Sucholutsky & Schonlau, 2021), accelerating model evaluation and architecture search (Zhao et al., 2020), and addressing privacy concerns (Dong et al., 2022).

The variety of computational approaches to the task of selecting informative examples for learners raises new questions for human teaching strategies as well. Using criteria such as diversity, joint representativeness, or a combination of the two, work in dataset distillation formalizes different possible strategies driving human pedagogical sample selection. For example, when selecting examples of the "mammal"



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

category, teachers might select diverse exemplars to highlight the extremes (e.g., bats and whales), prototypical exemplars to illustrate the central mass of the distribution (e.g., dogs and cats), representative exemplars to illustrate variation in non-extreme category members (e.g., otters and elephants), or some combination of these. These strategies may further shift with quota: the number of examples teachers can choose to represent a category. Yet, existing accounts of human teaching have not systematically evaluated which objectives best characterize exemplar selection.

## Our Contributions

In this work, we aim to bridge this gap by leveraging tools from machine learning and dataset distillation to computationally model how humans select exemplars to teach others. We embed datasets using pretrained neural network feature representations and formalize multiple candidate selection strategies that differ in how they balance concepts such as representativeness of the overall dataset, prototypicality, and diversity. These strategies define concrete metrics for choosing a subset of exemplars from a larger set. We then test adult human participants on the same selection task and evaluate which computational metrics best capture human behavior.

By grounding models of human exemplar selection in learned representations and principled subset selection objectives, our approach offers a new way to connect machine learning methodologies of dataset distillation with cognitive theories of teaching and information transmission.

## Methods

### Stimuli

We used visual stimuli created by Havy and Waxman (2016), including members of 3 novel categories (labeled as "daxes", "veps", and "bems"). In each category, members lie on a one-dimensional spectrum between two distinct creatures, continuously morphing various attributes such as color, feature, shape, and size. Each member is assigned a *scale value* (0 – 100) that quantifies where on the spectrum it lies. These categories have previously been shown to be learnable by 2-year-old children (LaTourrette & Waxman, 2022).

We then selected a range of exemplars from each category, ensuring that each category varied smoothly across exemplars and categories were roughly aligned in their within-category similarity. As shown in Figure 1, the cosine similarity between each exemplar and the category midpoint peaks at the center of the scale and decreases smoothly and symmetrically as exemplars become more extreme. Thus, the midpoint is genuinely prototypical with respect to the chosen range and deviations toward either endpoint are balanced. As the category similarities computed by the ResNet varied smoothly and equally for these ranges and choice of midpoint, we fixed them to these values for our experiments.

### Neural Network-based Distillation

To characterize how different neural network architectures prioritize representativeness, prototypicality, and diversity, we select exemplar subsets based on alignment in learned feature representations by a pretrained neural network. In particular, given $N$ images $x_1, \ldots, x_N$ from a novel object category, we extract feature representations using pretrained convolutional and transformer-based vision models. Specifically, we consider a ResNet-50 (He et al., 2016) and a ViT-B/16 (Dosovitskiy, 2020) architecture, each pretrained on ImageNet. We chose these architectures because they are well-established vision models that differ in their inductive biases and representational geometry. ResNet-50 emphasizes local, convolutional feature composition, while ViT-B/16 is transformer-based and relies on global self-attention, allowing us to examine whether different representational assumptions lead to different predictions about exemplar selection. All feature vectors are normalized to have unit norm so that inner products correspond to cosine similarity. Let $f_i$ denote the normalized feature vector associated to the $i$-th image $x_i$. We define the cosine similarity between $x_i$ and $x_j$ as $\mathrm{sim}(i,j) = \langle f_i, f_j \rangle = f_i^T f_j$ along with their cosine distance $\mathrm{dist}(i,j) = 1 - \mathrm{sim}(i,j)$.

Leveraging these representations, we will choose a subset of exemplars with quota $M$ that optimizes a specified objective. Each objective aims to mathematically formalize different strategies for choosing exemplars.

**Prototypicality** To identify prototypical exemplars, we rank individual datapoints by their average similarity to the dataset. Given a quota of $M$ exemplars, we then select the top $M$ members according to this score. This criterion identifies stimuli that are considered the most typical based on the neural network's feature representation (e.g., for mammals, dogs, cats, and cows would score higher than whales or bats).

**Representativity** To identify exemplar sets that are jointly representative of the dataset, we use a facility location criterion (Mirchandani & Francis, 1990; Wei et al., 2015) to find a subset $S \subset \{1, \ldots, N\}$ of size $|S| = M$ that is most similar on average across the dataset:

$$\max_{|S|=M} \text{Representativity}(S) := \sum_{i=1}^{N} \max_{j \in S} \mathrm{sim}(i,j). \qquad (1)$$

This objective encourages selection of exemplars such that every data point is close to at least one selected exemplar, which implicitly discourages redundancy within the selected exemplars (e.g., it might favor selecting otters and elephants as examples of "mammal," rather than the more prototypical dogs and cats). Notice that, unlike the prototypicality criterion, this criterion evaluates representativeness at the level of the subset rather than the individual exemplar.

**Diversity** To measure diversity independently of representativeness, we select exemplars via

$$\max_{|S|=M} \text{Diversity}(S) := \sum_{i<j, i,j \in S} \mathrm{dist}(i,j). \qquad (2)$$

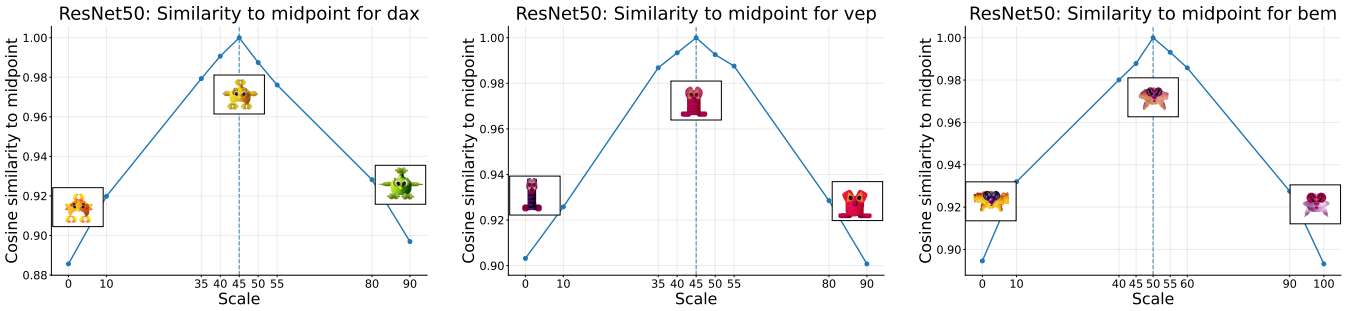


Figure 1: For each stimulus category (dax, vep, bem), we compute the cosine similarity between the ResNet feature representation of each member and the representation of the midpoint stimulus (dashed vertical line). We also visualize the left and right endpoints along with the midpoint.

This criterion favors subsets whose members are mutually far apart in the network's representation space, reflecting the network's notion of perceptual or semantic diversity. This generally results in better coverage of the category extremes (e.g., selecting whales and bats as "mammals"). We only use this criterion in the case when $M \geq 2$, since there is no notion of mutual diversity with only a single exemplar.

**Combining objectives** Using the above criteria, one can create hybrid objectives that also combine various attributes. For example, to study trade-offs between representativeness and diversity, we consider a combined objective

$$\max_{|S|=M} \text{Representativity}(S) + \text{Diversity}(S). \tag{3}$$

This objective aims to balance selecting exemplars that jointly represent the dataset and selecting exemplars that are maximally distinct from one another. Similar to the Diversity criterion, this Combination is only considered when $M \geq 2$.

Together, these selection criteria allow us to disentangle multiple notions of exemplar quality in neural network representations: individual typicality (prototypical exemplars), joint representativeness, and mutual diversity. By applying each criterion to features extracted from different architectures, we can directly compare how neural networks differ in what they consider representative, diverse, or extreme within the same stimulus set.

**Dependent measures** We assess the qualities of exemplars selected by both humans and machines through several measures. The first is a *prototypicality score*, which computes the average absolute distance of the exemplars chosen to the prototypical member, given by the midpoint. Recall that each member of a category is given a scale value. Given the midpoint's scale value $m_s$, we compute the prototypicality score of a subset of exemplars with scale values $s_1, \ldots, s_M$ by

$$\text{prototypicality score} = \frac{1}{M} \sum_{i=1}^{M} \frac{|s_i - m_s|}{m_s}.$$

We additionally consider a *diversity score* by computing the maximum pairwise distance between exemplars with scale values $s_1, \ldots, s_M$ via

$$\text{diversity score} = \begin{cases} \max_{i \neq j} \frac{|s_i - s_j|}{90} & \text{if category is dax or vep,} \\ \max_{i \neq j} \frac{|s_i - s_j|}{100} & \text{if category is bem.} \end{cases}$$

Note that we normalize prototypicality by the category midpoint and diversity by the category range so that they are on the same scale and lie between 0 and 1.

## Human Participants

**Participants** Twenty-four adult participants were recruited from a university participant pool for course credit.

**Design** The study was administered in Qualtrics and consisted of three blocks. Each block contained three trials, each featuring a different novel category (see Figure 1) and exemplar quota. Thus, within each block, participants viewed all three categories, each assigned a different exemplar quota (1, 2, or 3). Across blocks, each category appeared once in each quota condition. The order of the questions within each block and the presentation of the blocks were randomized.

**Procedure** Participants were introduced to the task with a cover story in which they were asked to imagine they were astrobiologists studying aliens from a distant planet. They were told that some aliens had already been reliably identified as members of a given species, and their task was to teach new trainees how to correctly identify each species.

Participants then completed a series of trials corresponding to the different exemplar-quota conditions. In each trial, they were told, "These are all [species name, e.g., daxes]. You can choose only [exemplar quota, e.g., one] to teach someone about [daxes]. Which [one(s)] would you pick?" Participants then selected the exemplar(s) they would use for teaching.

## Human Results

We first report results for human participants before modeling comparisons. For each participant, we computed a prototypicality score for all exemplar quotas, and the diversity score for 2- and 3-exemplar quotas. We conducted planned analyses to examine whether (1) participants' prototypicality scores increased as the number of exemplars available to

select increased, (2) diversity differed between the 2- and 3-exemplar conditions, and (3) participants' prototypicality and diversity scores exceeded chance in each exemplar condition.

**Prototypicality Score** We first computed chance performance by taking the average distance from each stimulus to the category midpoint under uniform random selection with no sensitivity to category structure. We then normalized the chance prototypicality score by dividing it by the category midpoint's scale value. This yielded a chance prototypicality score of 0.466 for bems, and 0.469 for veps and daxes. Figure 2 shows prototypicality scores by quota.

We fit a linear mixed-effects model on chance-centered and normalized prototypicality scores with fixed effects for the exemplar quota condition (treated as a factor with three levels) and random intercepts and slopes for quota by participant. In this measure, higher values indicate greater deviation from the midpoint and thus less prototypical selections. We found a significant effect of exemplar condition (ref = 1-exemplar condition), $b = 0.337$, $SE = .064$, $p < .001$; $b = .325$, $SE = .073$, $p < .001$ for the 2- and 3-exemplar conditions, respectively–indicating that participants deviated more from the midpoint as they made more selections.

We then constructed an intercept-only model for each condition. In the 1-exemplar condition, selections did not differ reliably from chance (intercept = -.106, $SE = .072$, $p = .15$). Although the most frequently selected exemplar was the category midpoint (26.4% of trials), the next most frequent selections were the two end points (12.5% of trials each). This suggests heterogeneity in participants' sample selection strategies: while many favored a central exemplar, a subset of participants preferred boundary cases. In contrast, prototypicality scores were significantly higher than chance when the participants were allowed to select two ($b = .231$, $SE = 0.04$, $p < .001$) or three exemplars ($b = .219$, $SE = .013$, $p < .001$). This indicates that given the opportunity to select multiple exemplars, adults chose exemplars significantly further from the midpoint than when they were allowed to choose only one exemplar. This preference for selecting typical exemplars when given only one exemplar, but exemplars further from the midpoint when given 2 or 3, suggests adults flexibly adapt their sampling strategy across quota constraints.

**Diversity Score** We estimated chance-level diversity in each category and exemplar quota by computing and normalizing the average expected diversity for all possible combinations of exemplars. This yielded an average maximum diversity of .38 for 2-exemplar quotas and .57 for 3-exemplar quotas (as expected, more selections led to greater average diversity). To facilitate comparisons against chance, we subtracted the normalized chance score from the observed score, and ran intercept-only linear mixed-effects models for each quota condition with random intercepts for participants.

In the 2-exemplar condition, the mean chance-centered diversity was significantly greater than zero ($b = .225$, $SE = .05$, $p = .0001$), suggesting that selections were on average 22.5 percentage points farther apart on the 0-100% scale than

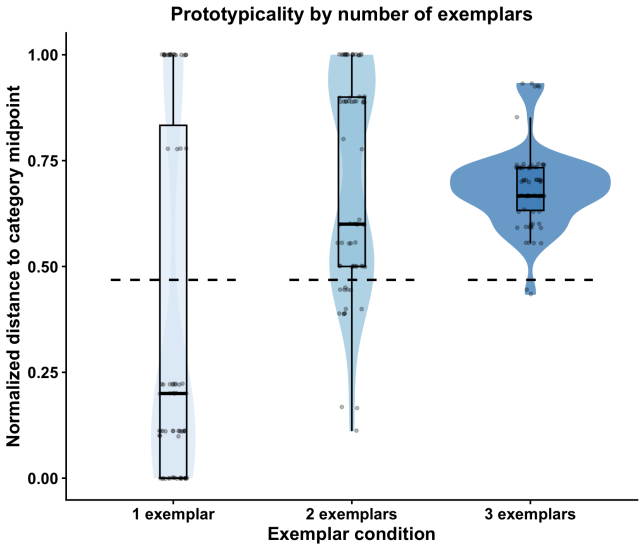


Figure 2: Prototypicality of selected exemplars by condition. Dotted horizontal line indicates chance performance. As the quota for the number of exemplars increases, the prototypicality score tends to increase, meaning that the chosen exemplars become less typical and examples at the endpoints are chosen more frequently.

expected by chance. In the 3-exemplar condition, mean diversity was even greater ($b = .338$, $SE = .023$, $p < .0001$), with participants' selections averaging 33.8 percentage points more diverse than chance. This is especially notable as the chance-based diversity score was higher in the 3-exemplar condition, yet participants still outperformed it.

We then constructed a linear mixed-effects model predicting chance-centered diversity scores with exemplar quota (2 vs. 3) as a fixed effect and random intercepts and slopes for quota by participant. We observed a significant effect of condition: the diversity score in the 3-exemplar condition was higher than in the 2-exemplar, $b = .112$, $SE = .045$, $p = .019$. Thus, when allowed to select three exemplars, instead of two, adults chose even more distinct exemplars.

## Modeling Results

We now discuss the modeling results using the neural network features. First, we showcase exemplar selections for each architecture and choice criterion. Table 1 reports the normalized selections of exemplars by architecture, separated by choice criterion and shown alongside the most popular human selection. Across criteria and categories, the models exhibit systematic differences in which exemplars are selected.

**Comparison with human data** Next, we consider if the choices made by the neural network capture the choices made by human participants. To measure accuracy, we compute prototypicality and diversity scores for each participant and architecture–criterion on each category/quota pairing (e.g., "bems"/quota 3), and then calculate the absolute difference

| Quota | ResNet | | | | ViT | | | | Human |
|---|---|---|---|---|---|---|---|---|---|
| | Prototypicality | Representativity | Diversity | Combination | Prototypicality | Representativity | Diversity | Combination | |
| 1 | 0.5 | 0.5 | – | – | 0.5 | 0.5 | – | – | 0.5 |
| 2 | 0.5, 0.6 | 0.3, 0.8 | 0, 1 | 0.2, 1 | 0.4, 0.5 | 0.3, 0.8 | 0.2, 1 | 0.3, 0.8 | 0, 1 |
| 3 | 0.4, 0.5, 0.6 | 0, 0.5, 0.9 | 0, 0.5, 1 | 0, 0.5, 1 | 0.4, 0.5, 0.6 | 0, 0.5, 0.9 | 0, 0.5, 1 | 0, 0.5, 1 | 0, 0.5, 1 |

Table 1: Normalized model exemplar locations by quota. For each quota $M$ and selection criterion, we report the model's mean normalized exemplar locations (averaged across categories after normalizing scales to $[0, 1]$). In the Human column, we report the most popular choice for each quota, also normalized and averaged across categories. No selection is made for the Diversity and Combination criteria when $M = 1$.

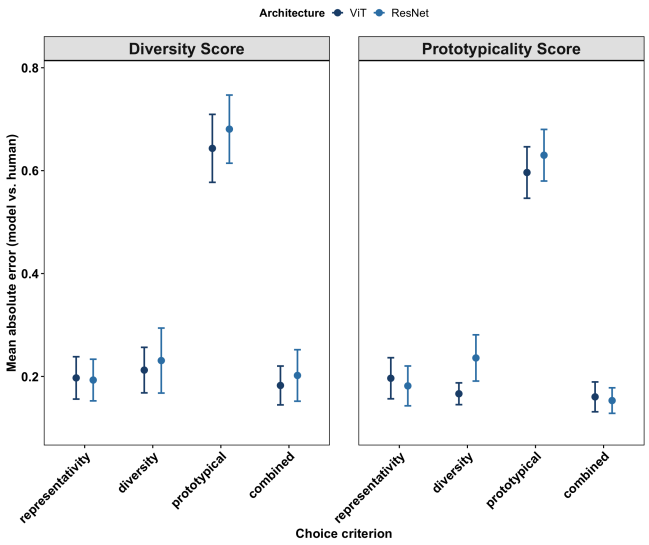


Figure 3: The mean absolute error of the prototypicality score and diversity score between the neural network model and human participants, collapsed across all participants, exemplar quotas 2 and 3, and categories. Error bars are 95% CI around the mean participant error.

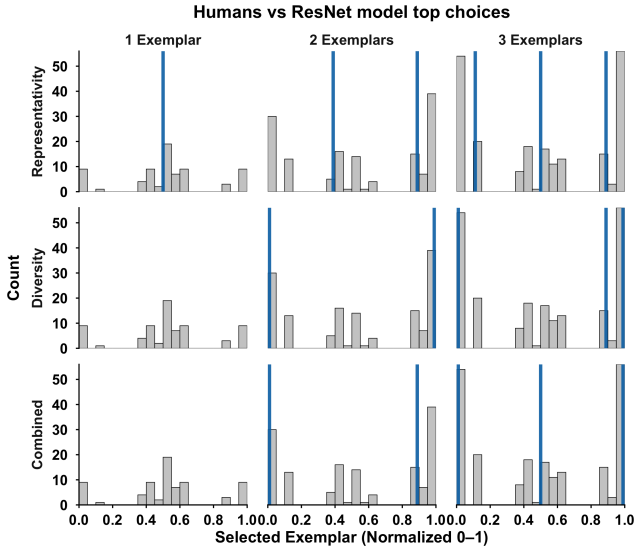


Figure 4: ResNet exemplar selection by choice criterion for sample category (blue lines) overlaid on human selections across categories.

between them, averaging this error across participants. We conduct this analysis when 1) aggregating over the 2- and 3-exemplar conditions and 2) via a more fine-grained analysis comparing model performance for each quota. Since the diversity criterion is only used in cases where $M \geq 2$, and the prototypicality and representativity criterions make identical predictions, we exclude quota 1 from the aggregate analysis.

**Aggregating across exemplar conditions** We report the mean absolute error between the neural network predictions and human participants for both the prototypicality score and diversity score aggregated across exemplar quotas 2 and 3 in Figure 3. Both the ViT and ResNet achieve the lowest mean prototypicality error (0.160 and 0.153, respectively) under a combined representativity-diversity criterion (Eq. (3)). For the diversity score, the ViT combined criterion achieves the lowest error (0.182), improving over simple representativity (0.197) or diversity (0.212), while the representativity criterion for the ResNet achieved the lowest error (0.193). Note also that for both models, the diversity criterion (Eq. (2)) is outperformed by the representativity and combined models when comparing diversity scores, which indicates maximizing mutual dissimilarity among exemplars alone does not capture human diversity judgments well. Averaging across measures, the combined criterion achieved the lowest error for both the ResNet (0.178) and the ViT (0.171), suggesting that human teaching behavior involves a balance between representativeness and diversity.

**Fine-grained condition analysis** Table 2 provides a fine-grained comparison of model-human alignment across exemplar quotas and measures. Because the stimuli were selected and validated using ResNet feature representations, we focus primarily on the ResNet comparisons, while treating the ViT results as a comparison across representational architectures. Across conditions, selection strategies based on representativity and a combination of representativity and diversity generally outperform strategies that optimize prototypicality or diversity alone. For the ResNet, the representativity criterion achieves the lowest average error across measures at the 2-exemplar condition, and the combined criterion achieves the lowest average error for the 3-exemplar condition. This shift

| Quota | ResNet Prototypicality | ResNet Representativity | ResNet Diversity | ResNet Combination | ViT Prototypicality | ViT Representativity | ViT Diversity | ViT Combination |
|---|---|---|---|---|---|---|---|---|
| **Prototypicality score** | | | | | | | | |
| 1 | <u>0.362</u> | <u>0.362</u> | – | – | **0.304** | **0.304** | – | – |
| 2 | 0.647 | 0.245 | 0.284 | **0.243** | 0.628 | 0.268 | 0.262 | <u>0.250</u> |
| 3 | 0.615 | 0.121 | 0.187 | **0.063** | 0.567 | 0.128 | <u>0.072</u> | 0.074 |
| **Diversity score** | | | | | | | | |
| 2 | 0.556 | **0.257** | 0.374 | 0.318 | 0.538 | <u>0.265</u> | 0.325 | 0.267 |
| 3 | 0.805 | 0.132 | **0.087** | **0.087** | 0.750 | 0.132 | <u>0.101</u> | <u>0.101</u> |
| **Average of prototypicality and diversity scores** | | | | | | | | |
| 2 | 0.602 | **0.251** | 0.329 | 0.281 | 0.583 | 0.267 | 0.294 | <u>0.259</u> |
| 3 | 0.710 | 0.127 | 0.137 | **0.075** | 0.659 | 0.130 | <u>0.087</u> | 0.088 |

Table 2: **Model–human alignment across exemplar quotas and selection criteria.** We report mean absolute error (MAE) for prototypicality (top), diversity (middle), and their average (bottom). Bold values indicate the lowest error within each row; underlined values indicate the second lowest.

likely reflects a change in human selection strategy with an increase in quota. Participants covered more of the category at Q3 than at Q2, such that incorporating a diversity component in addition to representativity better captures their behavior.

To make more direct comparisons with human choices, Figure 4 provides a qualitative comparison between human exemplar selections and the exemplars chosen by the ResNet under different choice criteria and exemplar quotas. The diversity criterion tends to focus on extreme endpoints of the morph continuum, while the representativity and combined criterion incorporate choices near the midpoint along with choosing less extreme endpoints (see, e.g., the representativity choices in the 2- and 3- exemplar conditions).

## Discussion

The present work examined how humans select exemplars to teach others and whether principled selection objectives from machine learning can serve as useful computational models of this behavior. Our results suggest that human teaching behavior is best characterized not by a single objective, but by a structured balance between representativeness and diversity that shifts with resource constraints. This pattern suggests that human teaching aligns with preferences in human learners, mirroring findings in category-based induction where learners draw stronger inferences from representative and diverse samples than from atypical or redundant ones (Heit, 2000; Osherson et al., 1990; Rhodes et al., 2008). That is, human teachers appear to select exemplars in ways that match learners' inductive sensitivities.

Adults' teaching decisions shifted with quota constraints, with selections becoming less prototypical and more diverse as exemplar quota increased. The shift in adults' teaching decisions was captured by the ResNet model selections, with the advantage of the representativity criterion (selecting exemplars that collectively summarize the dataset) evident at quota 2 and the combined objective (combination of representativity and diversity) winning out at quota 3. For the ViT, the combined criterion achieved the lowest error when averaging across quotas and measures, suggesting that a combination of representativity and diversity captures human teaching, even when the feature space is not the one used to validate the stimuli. Overall, adults appear to structure selections to jointly represent the distribution while adjusting their selections as quota changes. Our finding supports the idea that human teachers may implicitly trade off multiple pedagogical goals, rather than optimizing a single criterion.

This work highlights the value of using subset selection objectives as tools for modeling human teaching. However, the neural network representations we use remain complex and difficult to interpret directly. It would be of interest to develop smaller or more interpretable representational models that can capture human exemplar selection behavior while offering greater explanatory insight. Further, since we used a simplified stimulus set for experimental control, the generalizability of the present findings is limited. Future work might explore how the same processes work for real-life categories that vary along multiple dimensions. Another promising direction concerns individual differences in exemplar selection. As noted above, there is some heterogeneity across teachers (especially in 1-exemplar contexts); subsequent work might model individual teachers and test how strategies shift based on context. Future work might also consider how such strategies emerge over development, building on findings that younger children prioritize category typicality over diversity (Rhodes et al., 2008). Finally, the present study focused on the teaching side, yet teaching and learning are two interrelated processes. It would be of interest to examine whether certain teaching criteria (e.g., representativity or combined objectives) result in better category learning. This would allow future work to test not only which objectives describe human teaching, but also which objectives produce the most effective learning.

## References

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *The Quarterly*

*Journal of Experimental Psychology Section A*, *50*(3), 586–606.

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1149–1157.

Dong, T., Zhao, B., & Lyu, L. (2022). Privacy for free: How does dataset condensation help privacy? *International Conference on Machine Learning*, 5378–5396.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Har-Peled, S., & Kushal, A. (2007). Smaller coresets for k-median and k-means clustering. *Discrete and Computational Geometry*, *37*(1), 3–19.

Havy, M., & Waxman, S. R. (2016). Naming influences 9-month-olds' identification of discrete categories along a perceptual continuum. *Cognition*, *156*, 41–51.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic bulletin & review*, *7*(4), 569–592.

LaTourrette, A., & Waxman, S. R. (2022). Sparse labels, no problems: Infant categorization under challenging conditions. *Child development*, *93*(6), 1903–1911.

Lei, S., & Tao, D. (2023). A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(1), 17–32.

Li, J., Chen, P., Yu, S., Liu, S., & Jia, J. (2023). Bal: Balancing diversity and novelty for active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(5), 3653–3664.

Mirchandani, P. B., & Francis, R. L. (1990). *Discrete location theory*.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological review*, *90*(4), 339.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*, *34*(2), 133–143.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Qiu, F. W., Park, J., Vite, A., Patall, E., & Moll, H. (2025). Children's selective teaching and informing: A meta-analysis. *Developmental Science*, *28*(1), e13576.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, *108*(2), 543–556.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science*, *13*(3), 421–429.

Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. *International Conference on Learning Representations*.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Sucholutsky, I., & Schonlau, M. (2021). Secdd: Efficient and secure method for remotely training neural networks (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(18), 15897–15898.

Tsang, I. W., Kwok, J. T., Cristianini, N., et al. (2005). Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, *6*(4).

Wang, T., Zhu, J.-Y., Torralba, A., & Efros, A. A. (2018). Dataset distillation. *arXiv preprint arXiv:1811.10959*.

Wei, K., Iyer, R., & Bilmes, J. (2015). Submodularity in data subset selection and active learning. *International conference on machine learning*, 1954–1963.

Zhao, B., Mopuri, K. R., & Bilen, H. (2020). Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*.