Evaluating the Utility of Conformal Prediction Sets for AI-Advised Image Labeling

Dongping Zhang
dzhang@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Angelos Chatzimparmpas
angelos.chatzimparmpas@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Negar Kamali
negar.kamali@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Jessica Hullman
jhullman@northwestern.edu
Northwestern University
Evanston, Illinois, USA

ABSTRACT

As deep neural networks are more commonly deployed in high-stakes domains, their black-box nature makes uncertainty quantification challenging. We investigate the effects of presenting conformal prediction sets—a distribution-free class of methods for generating prediction sets with specified coverage—to express uncertainty in AI-advised decision-making. Through a large online experiment, we compare the utility of conformal prediction sets to displays of Top-1 and Top-k predictions for AI-advised image labeling. In a pre-registered analysis, we find that the utility of prediction sets for accuracy varies with the difficulty of the task: while they result in accuracy on par with or less than Top-1 and Top-k displays for easy images, prediction sets excel at assisting humans in labeling out-of-distribution (OOD) images, especially when the set size is small. Our results empirically pinpoint practical challenges of conformal prediction sets and provide implications on how to incorporate them for real-world decision-making.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; **Visualization design and evaluation methods**; **Empirical studies in visualization**.

KEYWORDS

conformal prediction, image labeling, semi-supervised learning, comparative user experiment

1 INTRODUCTION

As machine learning models such as deep neural networks (NNs) achieve impressive predictive performance, AI-advised decision-making has become more commonplace in various situations where a human must make a decision, from high-stakes domains like medicine (e.g., [50]) or autonomous vehicle navigation (e.g., [16]) to online tasks like content moderation (e.g., [35]) or image labeling (e.g., [60]). Ideally, having access to a predictive model can help humans make more informed decisions, surpassing the capabilities of either human judgment or AI in isolation, a synergy known as *complementary performance* [6, 25]. Whenever an AI system understands certain regions of the feature space better, the human counterpart can benefit by considering the model predictions. For example, in AI-advised image labeling, which we study in this paper, human labelers may benefit from viewing predictions from a model trained on previously labeled examples.

To help humans know when to trust model predictions, presenting information about the probability that the model's prediction is correct can, in principle, be useful. For example, a decision-maker could be provided with the softmax pseudo-probability produced in the final layer of NNs for classification tasks. However, NN predictions are known to be overconfident at times [22], and heuristic measures of uncertainty are often poorly calibrated. Consequently, decision-making relying on NN predictions with the highest softmax scores (e.g., Top-1 or Top-k) can fail to account for prediction error, leading to worse decisions.

Given these challenges, conformal prediction [3, 52, 61] has emerged as an alternative solution that can rigorously quantify the prediction uncertainty for NNs through the use of *prediction sets*. Instead of presenting unreliable softmax values or other heuristics for uncertainty, the prediction set is a type of confidence interval consisting of a set of predictions with a *coverage guarantee*—the true class is, on average, captured within such sets with a user-specified probability (e.g., 95%). The conformal framework is compatible with NNs because the derived sets are *distribution-free*—they are model-agnostic, work with finite samples, and offer non-asymptotic guarantees without distributional assumptions [4].

One challenge is that while conformal prediction sets express prediction uncertainty with guarantees that status quo presentations like Top-k predictions with softmax cannot, the conformal *coverage guarantee* can lead to large sets embodying high uncertainty for instances the model perceives as difficult [3, 5]. Compared with

Top-k predictions, such sets may increase cognitive load, rendering uncertainty quantification less effective [65], thus compromising predictive performance. To understand how conformal prediction sets support human decisions aided by model predictions in an image labeling task, we contribute a large online repeated measure experiment ($n = 600$) that compares the accuracy achieved in AI-advised image labeling using conformal prediction sets with that achieved with no predictions or Top-k predictions. Participants in our study were tasked with labeling images from ILSVRC 2012 [13] that varied in difficulty and encompassed both image stimuli that were in-distribution and out-of-distribution (OOD). This allowed us to evaluate the utility of prediction sets against Top-k presentation alternatives in scenarios where the model's predictions were mostly accurate versus when the presence of "unknown unknowns" makes predictions more error-prone.

We evaluate decision-making using metrics such as *accuracy* and the *shortest path length*, which approximates the amount of deviation from the ground truth in the label space hierarchy. We use participants' elicited *willingness-to-pay* at the end of the experiment to compare the perceived value of predictions against the post-hoc inferred monetary benefits derived from prediction access.

We find that for in-distribution instances, prediction sets lead to reduced labeling accuracy compared to Top-k predictions. However, for OOD instances where model misspecification differentially affected the calibration of Top-k predictions, prediction sets improve accuracy regardless of the set size. We find that participants are willing to pay roughly equivalent amounts for each type of display; if anything, they undervalue prediction sets relative to other presentations. Our results advance understanding of the strengths and limitations of prediction sets for improving AI-advised decision-making and highlight decision-makers' tendency to over-rely on predictions, even when poorly calibrated, in an image labeling task.

2 BACKGROUND

2.1 Uncertainty Quantification Using Conformal Prediction

Prior work on Human-Computer Interaction (HCI) suggests that humans can make more informed decisions when uncertainty is effectively communicated (e.g., [29, 31, 33]), such as by presenting prediction uncertainty through static intervals [11, 12, 40, 57] or animations [30, 63]. However, uncertainty quantification for NNs remains challenging [2, 4] due to their black-box nature and unreliability of internal heuristic confidence values, such as softmax pseudo-probabilities [22, 58].

One approach to quantify prediction uncertainty in classification tasks is using Bayesian NNs, where predictions depend on sampling from the posterior distributions of model parameters, carrying an inherent, mathematically grounded measure of uncertainty. However, Bayesian NNs heavily rely on distributional assumptions for their priors and can be computationally intensive to train. *Conformal prediction* [61] has emerged as a more flexible and efficient solution that does not require strong model or distributional assumptions. In classification settings, conformal prediction quantifies uncertainty through *prediction sets*. Similar to traditional confidence intervals, the derived sets contain predicted classes and achieve a *coverage guarantee* that, on average, the true class is captured by the set

with a user-specified probability. Formally, given a training set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is the feature vector and y_i is the corresponding label, the uncertainty set function, denoted $C(X)$, maps a feature vector x_i to a subset of $Y = \{1, \ldots, K\}$. This mapping ensures *coverage*[1] that, for any new instance x drawing from the same distribution as the training data, the probability $P(Y \in C(X)) \geq 1 - \alpha$ over the randomness in the calibration and test points, with an error rate of α [4].

If an uncertainty quantification approach is well-calibrated, we naturally expect it to reflect greater uncertainty on more difficult instances for the model (e.g., OOD instances). With prediction sets [3, 52], this manifests as *adaptiveness*: a conformal prediction set will be larger when the task is more difficult for the model. Adaptiveness is realized through the inductive conformal method, which quantifies the prediction uncertainty of a chosen classifier using a designated *calibration set* [45] separate from the training data set. This set contains instances that are independently and identically distributed (i.i.d.) to those on which the model is trained.

For each calibration instance, the predicted softmax scores are adjusted using Platt scaling [22, 43, 47] for better interpretability and further regularized to penalize unlikely classes [3]. These calibrated probabilities are then ranked in descending order, reflecting the classifier's confidence in each class. A conformity score is computed by summing these ranked probabilities up to and including the true class, capturing the model's accumulated confidence for the instance leading up to the correct classification. From the distribution of the conformity scores of all calibration instances, a threshold is computed by taking the α quantile after a finite sample correction for robustness. For a new and unknown instance, this calibration threshold decides the size of the prediction set by including all labels whose softmax scores exceed this calibration threshold. In our study, we opted for the Regularized Adaptive Prediction Sets (RAPS) algorithm [3] to ensure that the prediction sets of our experiment are as small as possible while achieving, on average, the desired coverage rate.

2.2 Effects of Presenting Uncertainty in AI-advised Decisions

Pairing humans with an AI system in complex decision-making scenarios can elevate performance beyond what either can achieve individually when their strengths are complementary [6, 23, 55]. In annotation tasks with large label spaces, AI models can improve accuracy and efficiency [37]. Particularly, in AI-advised image labeling, various tools exist for applying automation to enhance human annotators' speed and accuracy [53]. However, performance on AI-advised tasks can be influenced by how predictions are presented—specifically, whether uncertainty is effectively communicated or not. Effective communication of uncertainty can improve the perceived trustworthiness of predictions [8, 36, 64]. Presenting a model's accuracy is necessary to provide humans with a well-defined decision problem in many AI-advised decision settings [28], and can encourage analytical thinking and reduce over-reliance on predictions [48, 64]. However, uncertainty quantification does not guarantee improved decision-making. Cognitive biases, such as tendencies

---

[1]Recent work explores alternatives to marginal coverage, including class-conditional coverage [19] and relaxed versions of conditional coverage (e.g., [27]).

to under- or over-react to information relative to rational standards, play a major role in how uncertainty is interpreted (e.g., [1, 59]). Additionally, uncertainty quantification can improve trust under low cognitive load, but can diminish trust when uncertainty is presented ambiguously, demanding high cognitive load [65].

There is limited empirical evidence on how prediction sets influence trust and accuracy in AI-advised decision-making. One exception is Babbar et al. [5], who investigated humans' perceived utility and performance using RAPS versus Top-1 predictions for labeling images from CIFAR-100 [34]. They found that participants using RAPS reported higher trust and perceived utility than those using Top-1 predictions. However, their study was unlikely to have achieved sufficient power to detect differences in accuracy between the two groups due to the very small sample size (15 participants per group). They suggest that smaller sets might lead to better accuracy, but this observation was based on a comparison between sets generated by two different conformal methods. Recent work by Straitouri and Rodriguez [56] compared labeling performance between two decision support systems: one that only allowed participants to select a label value from the prediction set and another that allowed them to freely choose one of 16 labels. Their results suggest that when the model is well-calibrated, decision-making solely based on prediction sets tends to be more reliable and accurate because of the coverage guarantee.

Motivated by overlapping research questions to Babbar et al. [5], we conducted a much larger experiment to evaluate the utility of different uncertainty prediction displays for in-distribution and OOD stimuli, each systematically categorized by the difficulty and size of the associated prediction set. This enabled us to explore changes in performance by instance type. Contrary to the restricted agency in Straitouri and Rodriguez [56], our study allowed participants to freely navigate a large label space, so as to allow for the possibility that even inaccurate predictions may serve as useful clues that help participants deduce the correct answer.

3 ONLINE EXPERIMENT

3.1 Overview

We conducted a large mixed-design repeated measure experiment on Prolific. Following a pre-registered[2] analysis plan, we evaluated how different displays of prediction uncertainty impacted performance on an AI-advised labeling task, including relative to performance without access to predictions. Figure 1 provides an overview of our stimuli generation process, which systematically varied whether image instances were in-distribution or OOD and easy or difficult, as well as the size of conformal prediction sets.

Each participant in our study labeled a set of 16 images sampled from the ILSVRC 2012. We assigned participants to one of four *conditions* representing different *prediction displays*: (1) no predictions (baseline), (2) Top-1 prediction with softmax, (3) Top-10 predictions with softmax, or (4) the RAPS prediction set [3].

In real-world scenarios, the deployed model can encounter unexpected OOD instances where the distribution of input features differs from what was trained and calibrated. We varied whether images were in-distribution or OOD within subjects. Among the 16

stimuli, 4 were in distribution, while 12 were not, as we expected greater variation in participants' accuracy for OOD.

In our experiment, OOD instances notably lowered the model's Top-k prediction accuracy. Although the corrupted images based on the ILSVRC 2012 dataset that we used as OOD instances could compromise the conformal coverage guarantees, the adaptiveness of RAPS led to good coverage despite OOD. However, the number of labels included in the prediction sets became much larger when the instance was harder for the model. To evaluate the impact of set size on the effectiveness of prediction sets, we balanced set size within subject by categorizing images as smaller or larger based on the size of the prediction set generated for the image. This factor enabled us to examine the performance of the prediction set between size levels, as previous work suggests that a smaller set size is more preferable [3, 52] and may yield higher accuracy for in-distribution images [5]. We expected smaller set sizes to be more useful in general, as they could reduce cognitive load while maintaining the conformal coverage guarantee for in-distribution images. Larger sets embodying high uncertainty could still be useful for labeling when participants were highly uncertain of what the image showed and consequently heavily relied on prediction sets. In our setting, set size could also be viewed as a nuanced measure of difficulty for in-distribution stimuli because images that required a larger set size to achieve the desired coverage were intrinsically more difficult [3].

For both in-distribution and OOD stimuli, we balanced difficulty within subject by easy and hard instances, using the median cross-entropy loss of in-distribution images as a threshold to define these levels. The former enabled us to assess whether participants were inclined to accept predictions that expressed high confidence. The latter allowed us to examine whether participants were more likely to modify or reject low-confidence predictions and instead rely on their own intuition.

Our study evaluated participants' performance by their labeling *accuracy* and the *shortest path length* between the chosen label and the correct label in the WordNet hierarchy [15, 42]. After they had labeled all images, we asked them to report their *willingness-to-pay* for the prediction display they used (i.e., a value ranging from 0 to 8 USD, the maximum they could earn) if they were to repeat the study with a similar distribution of instances. Finally, we elicited the strategies employed to approach the tasks.

3.2 Task and Stimuli Generation

3.2.1 Overview of Stimuli Generation. To generate stimuli, we sourced 80 task images from ILSVRC 2012, evenly divided between in-distribution and OOD (i.e., 40 of each). For both in-distribution and OOD images, we further categorized each set of stimuli by difficulty (between easy and hard) and by set size (between smaller and larger). Note that because the range of observed set sizes differed depending on the difficulty of the instance, set sizes were defined relatively rather than absolutely. This categorization resulted in four distinct groups, each comprising ten images, for both in-distribution and OOD categories. Image stimuli within each group were selected to exemplify their respective categories, determined by the model's (described below) prediction confidence and the size of the prediction set derived.

---

[2]Pre-registration: https://aspredicted.org/6gh27.pdf

1. Impose covariate shifts

2. Categorize by difficulty and size

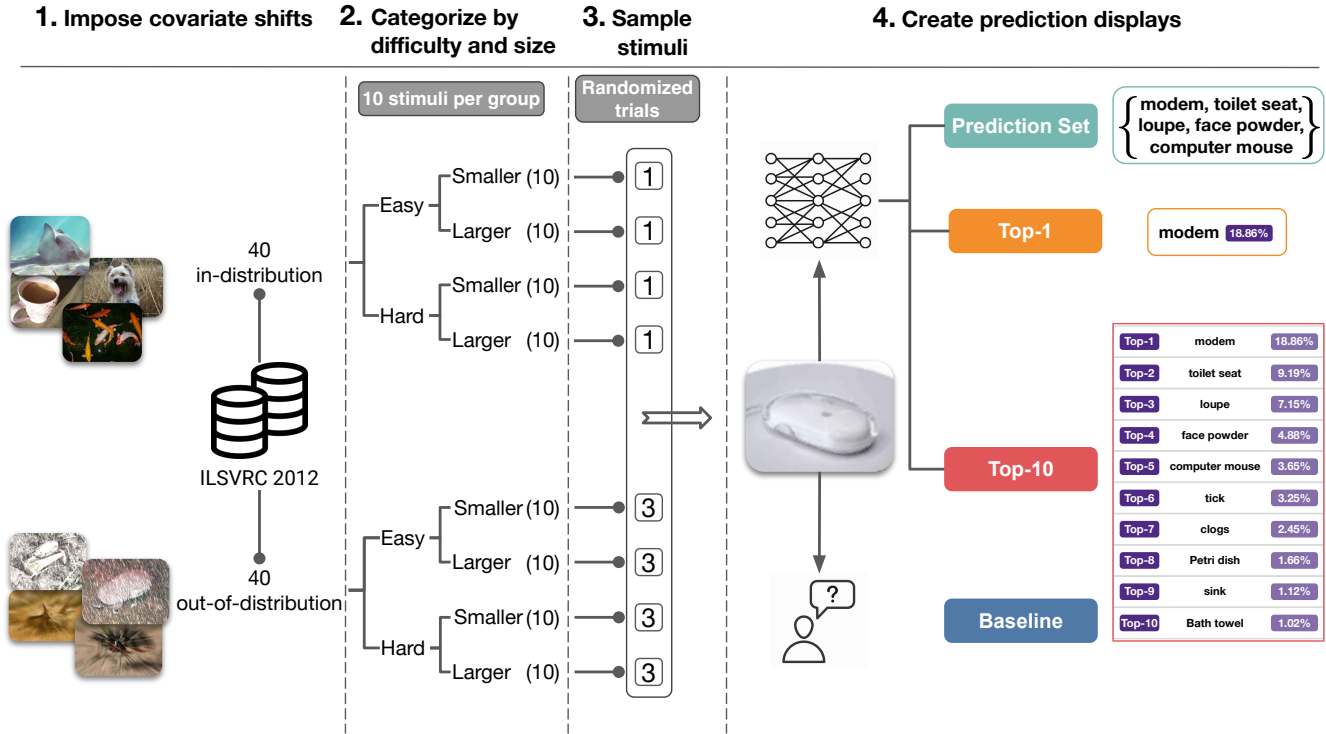**3. Sample stimuli**

4. Create prediction displays

Figure 1: Overview diagram of our key experimental manipulations. (1) Five different covariate shifts are imposed through synthetic image corruption to create five replications of the *conformal hold-out set*, each containing images that are OOD. (2) Images in each conformal hold-out set are categorized by the classifier's prediction confidence for difficulty and the size of the derived prediction set. Ten task images representative of the categories used to define each group are selected. (3) Participants label 16 task images sampled from 80 candidate images: four in-distribution and 12 OOD, balanced by difficulty and set size, presented in randomized order. Example task stimuli are shown in Figure 2. (4) Based on the conditions assigned, participants may complete labeling tasks without predictions (i.e., `baseline`) or with access to prediction displays that vary in the content provided by uncertainty quantification (i.e., `Top-1`, `Top-10`, or `RAPS`). Screenshots of the interface as seen by participants are presented in Figure 3.

3.2.2 Predictive Model for Image Classification. Given that underperforming models are rarely deployed in real-world scenarios, we chose a classifier designed to achieve high prediction accuracy. We used a pre-trained Wide Residual Network (WRN) [62] on ILSVRC 2012. The WRN is a Residual Network (ResNet) variant [24], and its primary architectural distinction is the widening factor in the channel width of the convolutional layers compared to a ResNet of the same depth. Specifically, we utilized a `wide_resnet101_2` model with 101 layers and a widening factor of 2, using PyTorch's `IMAGENET1K_V2` weights [46]. This model was well-calibrated for in-distribution instances, achieving a Top-1 accuracy of 82.5% and a Top-10 accuracy of 97.9% over all in-distribution images, as detailed in Table 6 of Appendix A. We used the WRN to create Top-1 and Top-10 predictions along with their corresponding softmax pseudo-probabilities to establish prediction displays for Top-k conditions.

3.2.3 Calibration Set. To produce prediction sets, we randomly sampled half of the 50,000 images from the ILSVRC 2012 validation set to create a *conformal calibration set*, which contained 25,000 in-distribution images that were unseen by the model. This calibration set was used to perform *inductive conformalization* by calibrating the model based on the conformity scores derived from all images in the *conformal calibration set*. This process enabled the creation of prediction sets for the remaining 25,000 images in the *conformal hold-out set*. Our prediction sets achieved ≈95% size-stratified conditional coverage ($\alpha = 0.05$), as shown in Table 1. This implies that, on average, there is a 95% probability that an image's true label class falls within the prediction set, conditioned on all set sizes.

Generating OOD images. To compare the utility of different prediction displays when the classifier is poorly calibrated, we utilized OOD images, which would lower the accuracy of Top-k predictions and increase the size of prediction sets. To create OOD images, we systematically transformed in-distribution images from our *conformal hold-out set* into OOD by applying *covariate shifts* [49] through synthetic image corruption, as described by Michaelis et al. [41]. Covariate shift, in the context of supervised learning where the joint distribution over images X and labels Y is $P(Y|X)P(X)$, refers to changes in $P(X)$, the feature probability, without affecting the classifier $P(Y|X)$.

Table 1: Coverage achieved by RAPS across various set size ranges with error rate $\alpha = 0.05$ for in-distribution stimuli in the *conformal hold-out set*.

Size	$\alpha = 0.05$	
	Count	Coverage
0 to 1	13,967	0.961
2 to 15	6,849	0.946
16 to 30	1,583	0.978
31 to 45	874	0.975
46 to 60	548	0.974
101 to 1,000	1,179	0.974

Table 2: The median RAPS size of images binarized by difficulty within each *conformal hold-out set*.

Hold-out Set	w/ Size = 1		w/o Size = 1	
	Easy	Hard	Easy	Hard
None	1	15	**5**	**19**
defocus blur	2	59	**8**	**60**
snow	1	41	**7**	**43**
frost	2	46	**8**	**50**
zoom blur	2	44	**10**	**47**
glass blur	2	63	**9**	**65**

After evaluating 15 types of corruption, we opted to use five corruptions that reduced WRN's Top-1 prediction accuracy the most. These corruptions were: (1) defocus blur, (2) snow, (3) frost, (4) zoom blur, and (5) glass blur, which are bolded in Table 6 of Appendix A. Given that the human visual system can be resilient to certain minor changes [26, 41], we applied these corruptions to maximum severity for all images in the *conformal hold-out set*, as defined by Michaelis et al. [41].

Table 6 of Appendix A indicates that our prediction sets achieve a 96% coverage rate for in-distribution images. However, the coverage guarantee for the OOD samples was compromised, as the images used during calibration and testing are no longer exchangeable (i.i.d.), according to Angelopoulos et al. [3].

Classifying stimuli by difficulty. To compare the effects of different prediction displays on images of varying difficulty, we categorized images by difficulty using the cross-entropy loss, which calculates how well the predicted probability distribution aligns with the ground truth distribution. Formally in Equation 1, let x be an input image; $p(x)$ denotes the true probability distribution of the image's label in one-hot encoding, while $q(x)$ is our classifier's predicted distribution.

$$H(p, q) = - \sum p(x) \cdot \log(q(x)) \qquad (1)$$

Figure 9 of Appendix A displays the distribution of cross-entropy loss of all in-distribution image stimuli, with the median presented as an orange dotted line. Due to WRN's high predictive power, this distribution was highly right-skewed, which created a problem: if we used this median threshold to categorize difficulty, we would include many images in the easy group that the model was both confident and accurate in predictions, so the derived sets embodied little uncertainty, synonymous with Top-1 prediction. We thus computed a new median threshold (0.72) after excluding images with a set size of one such that the difficulty threshold, shown as a red solid line, was defined based on images with non-deterministic sets. Images with cross-entropy loss falling below or on this threshold were classified as easy, while those above were deemed hard.

Because the cross-entropy loss derived from the predictive distribution of OOD images can be misleading as a signal of task difficulty due to model misspecification [22, 58], we used the red median threshold from the in-distribution data to categorize the difficulty of OOD images.

Classifying stimuli by set size. Images within each difficulty group of the *conformal hold-out set* induce varied prediction set sizes. For instance, while images classified as easy typically had a smaller set size, some easy instances required a larger set size to achieve the desired coverage. To explore how set size influences the efficacy of prediction sets, we categorized the stimuli within each difficulty group by their set size. After grouping images in each conformal hold-out set by easy and hard, we further distinguished them based on *relative* set size, as determined by RAPS, into smaller or larger categories, based on the median of the group. However, as shown in Table 2, the predictive power of WRN consistently demonstrated high confidence in its predictions for many easy images with a set size of one, which significantly lowered the median size threshold. To introduce more variance in set size within the easy categories, we calculated the median threshold after excluding images with a set size of one (i.e., these were directly assigned to the smaller group). Images with a set size smaller than or equal to this adjusted size median were classified as smaller; if the opposite is true, they were classified as larger. The specific thresholds used for each difficulty category within the *conformal hold-out set* are outlined in Table 2, with the bold figures representing the applied thresholds.

Sampling task images based on defined categories. Upon categorizing all images in each *conformal hold-out* (i.e., one in-distribution and five OOD, totaling $6 \times 25,000$ images) by difficulty (easy and hard) and size (smaller and larger), our objective was to select a set of task images from which to sample image assignments for participants, so as to avoid results that overfit to a very small set of stimuli. Specifically, we targeted 80 task images from the 1,200 candidate images. Our goals were to ensure that: (1) the candidate images are representative of each group defined by in-distribution or OOD, difficulty, and size ($6 \times 2 \times 2$), and (2) task images avoid the known data anomalies in ILSVRC 2012 [14, 44].

Sampling images from groups, such as in-distribution, easy (difficulty), and smaller (set size), was challenging due to the prevalence of images with a set size of one for which prediction sets become nearly identical to the Top-1 prediction. To avoid oversampling sets of size one, (1) we temporarily set aside images with a set size of one, (2) we took images with sets larger than one whose set size falls near the median size in that group (i.e., 45th and 55th percentiles), and (3) we performed a post-hoc sampling correction to include images with a set size of one in proportion to their original presence in each group, as shown in Table 8 of Appendix A.

Table 3: Prediction accuracy, counts, and set size of the selected 80 task stimuli grouped by our experimental manipulations.

Shift	Difficulty	Size	Count	Coverage			Avg. Set Size
				Top-1	Top-10	RAPS	
in	easy	smaller	10	1	1	1	2.6
		larger	10	1	1	1	19.5
	hard	smaller	10	0.4	1	0.9	8.0
		larger	10	0.5	0.9	0.9	51.5
out	easy	smaller	10	1	1	1	**5**
		larger	10	1	1	1	**28.1**
	hard	smaller	10	0	0.6	0.9	**30.1**
		larger	10	0.1	0.2	0.9	**90.8**

Figure 2: We present four example stimuli (from a total of 10) for each combination of in-distribution or OOD, difficulty, and size categories. The rows differentiate between in-distribution and OOD, while the columns vary by difficulty and set size.

Images from ILSVRC 2012 are known to be noisy, with numerous data anomalies. To create the set of 80 task images, we sampled 50 *candidate* task images from each representative subset defined by in-distribution or OOD, difficulty, and size ($6 \times 2 \times 2 \times 50$) without replacement. We manually inspected the task images, starting from images with higher cross-entropy loss (hard), and omitted those with (1) an obviously wrong ground truth label, (2) a text overlay with the correct label, (3) no apparent focal object or many surrounding objects, (4) unusual dimensions, and (5) small multiples. Based on these rules, we selected 10 images from each of the 4 groups from the combination of [easy, hard] × [smaller, larger] that were in-distribution, and 2 images from each of the 4 groups with corruption. This process generated 80 task images in total, balanced by in-distribution or OOD, difficulty, and size.

Table 3 presents the Top-k accuracy, coverage, and average set size, grouped by our experimental factors, with example stimuli shown in Figure 2. Our stimuli exhibited the desired behavior, with the average prediction accuracy and set size varying in response to the manipulated difficulty and set size. Notably, while the stimuli we selected to achieve balance over our experimental manipulations resulted in close conformal coverage for OOD instances to the in-distribution case, anecdotally we found that OOD coverage tended to remain high (≈80%) over varying stimuli sets under our data generating model.

3.2.4 Reward and Bonus. Participants earned a base compensation of $4 for completing the study and received a $0.25 bonus for each correct label. With 16 trials, the maximum bonus was $4, bringing the total potential earnings to $8. We determined the task reward to
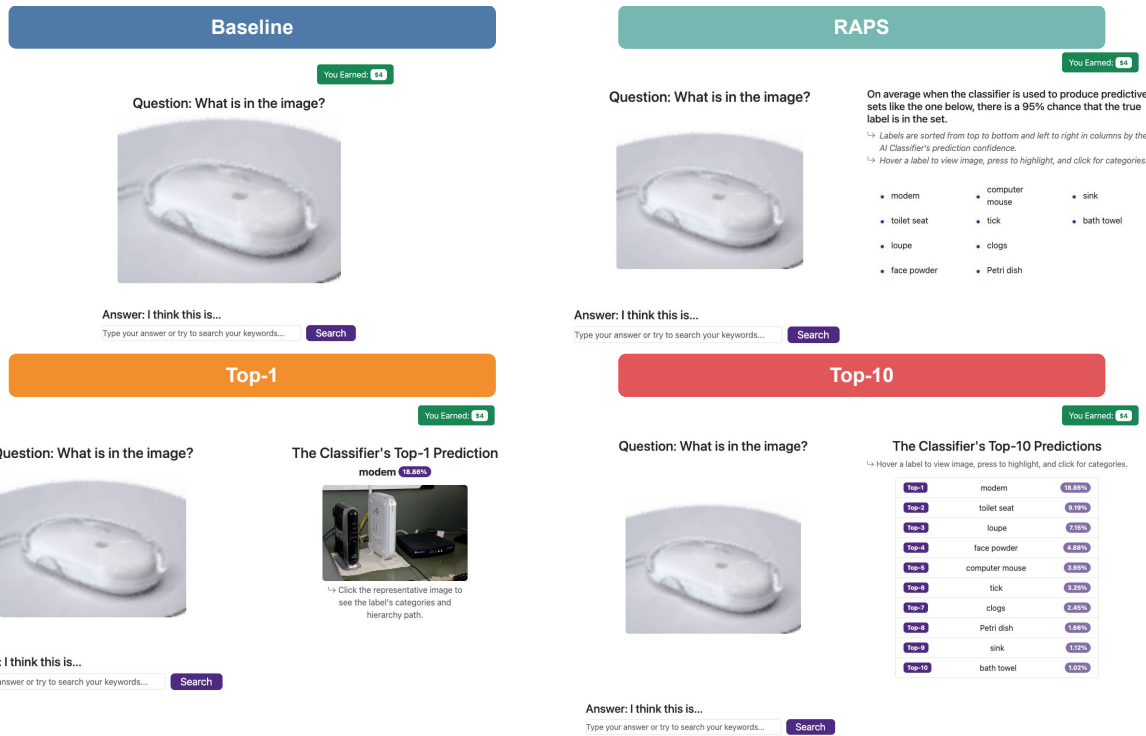
Figure 3: Screenshots of the interface participants used to complete the study by conditions (`baseline`, `Top-1`, `Top-10`, `RAPS`).

achieve the local minimum wage based on the assumption that most participants would take longer than the average time we observed in a small pilot of the experiment.

3.3 Experimental Interface

Our task interface, screenshots of which are presented in Figure 3, featured a two-column design: task images were on the left, and predictions were on the right, with a response field directly below the task image. In the baseline condition without predictions, the task image and the response field were centered. Participants could view their real-time earnings in the top-right corner of the interface.

To ensure that participants understood the meaning of each label, we selected label-representative images from the *conformal calibration set*. To find suitable label-representatives, we manually examined all images by labels and prioritized selection for those with small cross-entropy loss (i.e., easy), applying the same rules used when selecting task images. The label-representative images we used are included in the Supplemental Material.

3.3.1 Presenting Predictions. For the three treatment conditions that featured predictions, the Top-1 condition presented the classifier's Top-1 prediction and its softmax score on top of a representative image of the predicted label. Top-10 predictions with softmax scores were ordered in a table, while the prediction set was presented in an *n*-by-3 matrix sorted column-wise by the classifier's prediction confidence. Participants could hover their mouse cursor over any predicted label to view a label-representative image.

3.3.2 Label Space Search. To label task images, participants need to be aware of the label space. Because ILSVRC 2012 has a large label space of 1,000, cumbersome to present in a dropdown or checkbox format, we organized labels using a version of the WordNet hierarchy, which we truncated slightly to avoid label overlap. For example, "automobile" and "truck" are both parent nodes for "minivan" and since "truck" is also a type of "automobile", we retained "minivan" in the "automobile" category. We include this tree in the Supplemental Material.

If participants disagreed with the predictions or when no prediction was available (baseline), our interface enabled them to perform three types of search. As illustrated in Figure 4, they were (A) dropdown search, (B) keyword search, and (C) bottom-up search. In the dropdown search, a dropdown menu would appear as participants began typing into the response field, suggesting leaf labels that matched their input. Alternatively, in the keyword search, participants could type and then click a search button to search the hierarchy for specific keywords. Doing so would open an overlay window showing the label-space hierarchy network, where leaf nodes (i.e., denoting valid labels) would appear in a different color from the parent nodes. Clicking on a leaf node would populate the label in the response field. To ease navigation, the hierarchy network was collapsed by default except for the parent nodes (i.e., categories) and leaf nodes (i.e., labels) containing the search keywords with edges connecting them highlighted in orange. All other parent nodes could also be clicked to expand.

Participants could also perform a bottom-up search starting from the prediction display. Clicking on a predicted label would open the
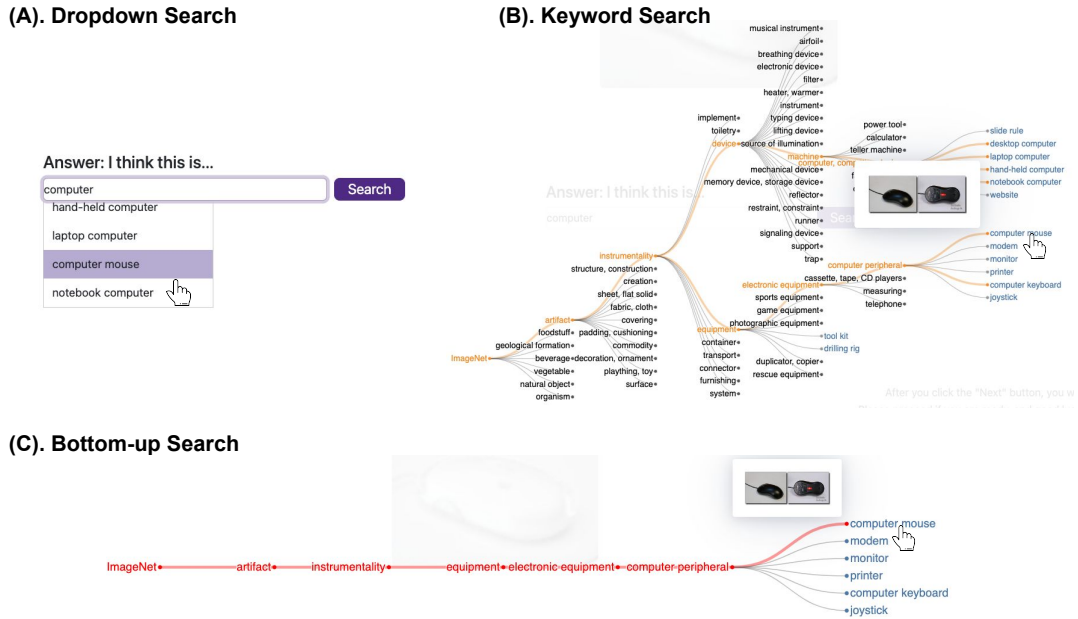
(A). Dropdown Search

(B). Keyword Search

(C). Bottom-up Search

Figure 4: Participants are provided with three search options to find their preferred choice. (A) *Dropdown search*: As participants type in the response field, a dropdown menu appears, exemplified by the entry "computer". (B) *Keyword search*: Participants can search their typed keywords in the WordNet hierarchy by clicking the "Search" button, which opens an overlay displaying a hierarchy network with the relevant network components highlighted. We provide an example of the rendered hierarchy network by searching for "computer"; (C) *Bottom-up search*: By clicking on a predicted label, such as "computer mouse", participants can see its path from the root node with categories on the path that can be clicked to expand for further exploration. Additionally, while exploring the hierarchy network, participants can hover over any leaf node to see label-representative images, as shown in (B) and (C).

search overlay. Instead of showing the entire hierarchy network, the overlay presented the path from the root node to the leaf node corresponding to the predicted label that the participant clicked on. Participants could explore adjacent labels under the same parent category or explore other parent nodes on the path toward the root.

When exploring the label space hierarchy, participants could hover over any leaf node to display its label-representative image. We illustrate these search actions through video demonstrations included in the Supplemental Material.

3.4 Experimental Procedure

Participants were directed to our study interface from Prolific. On the welcome page, they received a brief description of the study's purpose and the estimated time required for completion. Participants were instructed to complete the study in a single session using Google Chrome on a large-screen device. If participants agreed to the terms by clicking the "Start" button, they would be randomly assigned to a prediction display condition with task images sampled from each corresponding group, shuffled in a randomized order.

Participants then reviewed two instruction pages. The first introduced the interface layout and detailed the bonus mechanism. The second contained several short videos demonstrating how to explore predictions (except in the baseline condition) and navigate the label space. Participants had to watch all provided instructional

videos before proceeding to an example page where they could practice using the interface. Then, they completed 16 rounds of tasks. Upon finishing the 16 trials, participants were asked to state their willingness-to-pay to have access to the same style of prediction display they used in the experiment on a scale from 0 to 8 USD (i.e., the range of possible earnings from the experiment) if they were invited to undertake another 16 rounds of tasks with new images generated similarly to those they had experienced, for further reward and potential bonus. On the same page, participants could optionally describe their strategy to identify the correct label for each image, with or without the aid of predictions.

3.5 Participants

We recruited 600 participants from Prolific, divided equally among the four prediction displays, ensuring a gender-balanced sample. We set additional screening criteria so that all of our participants were (1) based in the USA, (2) had no vision impairment, (3) had no colorblindness, (4) spoke English as the first language, and (5) aged between 18 and 65. This led to 150 participants in each condition representing different prediction displays. We determined the target sample size using a non-parametric simulation of the experiment bootstrapped from a pilot study of 30 participants. By grouping trial-level responses according to all manipulated factors and bootstrapping these responses with replacement, we simulated

trial-level outcomes. We identified the target sample size (600) that resulted in 95% confidence intervals on accuracy with widths less than 10%.

3.6 Analysis Method

We pre-registered an analysis plan focusing on three response variables: (1) accuracy, (2) shortest path length, and (3) willingness-to-pay, and followed our pre-registered plan in all analyses reported below unless otherwise specified. We define accuracy (ACC) as binary, indicating whether a participant selected the correct label for the task. The shortest path (SP) length is a continuous variable representing the number of hops from the leaf node a participant selects to the leaf node of the ground truth label. SP length is a measure widely used to quantify semantic similarity [9, 51] between categories in a taxonomy. Although highly correlated with accuracy, this measure provides further information on how "incorrectness" varied by our experimental manipulations. Willingness-to-pay (WTP) ranges from $0 to $8, measuring how much they will pay to access the display.

We pre-registered Bayesian Linear Mixed-effect models for accuracy and SP length. Our model specifications are presented in Model 1 and Model 2. In this context, *trial* is numeric, indicating the trial number, whereas *condition* (4 levels: baseline, Top-1, Top-10, and prediction set), *shift* (2 levels: in-distribution and OOD), *difficulty* (2 levels: easy and hard), and *size* (2 levels: smaller and larger) are factors corresponding to our experimental manipulations. The term *ID* represents each participant's unique Prolific ID.

Model 1: Accuracy (ACC)

1: $acc \sim Bernoulli(p)$
2: $logit(p) = condition * shift * difficulty * size * trial + (trial|ID)$

For the accuracy model stated in Model 1, we use a Bernoulli distribution for the likelihood shown in Line 1, parameterized by p, which denotes the probability of a correct label. Line 2 presents the hierarchical logistic regression model. In this specification, we model the logit of p through interactions among all fixed effects and incorporate random intercepts and slopes for trial, grouped by participants' unique *ID*. This approach allows us to account for participants' baseline performance differences and variations in rates of change across conditions and trials.

Model 2: Shortest Path (SP) Length

1: $sp \sim ZeroInflatedNegBinomial(\mu, \theta, \psi)$
2: $log(\mu) = condition * shift * difficulty * size * trial + (trial|ID)$
3: $logit(\psi) = condition * shift * difficulty * size * trial$

For the SP length model stated in Model 2, we employ a Zero-inflated Negative Binomial (ZINB) model as the likelihood, parameterized by μ (expected SP length), θ (negative binomial dispersion), and ψ (zero-inflation probability), as shown in Line 1. The expected path length between the participants' responses and the truth label is modeled by the negative binomial component in Line 2. We apply a log-link function to μ, modeled by all fixed effects along with their

interactions with random intercepts and slopes for trials grouped by participant ID. We do not explicitly model the dispersion parameter for the NB component. However, we placed an exponential prior with a rate of 0.1 to moderate the degree of dispersion in our model. Given that our observed data had a maximum SP length of 27, this prior acts as a regularizing component, ensuring our model captures the dispersion of the observed data without predicting implausibly large SP lengths. Line 3 shows the zero-inflation component that models the proportion of zero path length (i.e., correct answer) not captured by the NB components. We apply a logit link function to ψ and model it by all fixed effects along with their interactions.

We deviated from our pre-registered model specifications only in setting the priors. We initially aimed to set weakly informative priors, but several were too narrow to allow model convergence. We therefore widened the priors to be even less informative. Full details are available in the Supplemental Material.

We followed a standard Bayesian workflow [17] to check if both models fit. Detailed model diagnostics are included in the Supplemental Material. We use the models to generate predictions of the respective response variable based on 1,000 draws from the posterior distribution of the models' fixed effects for each unique combination of our experimental manipulations while holding the trial effect constant at the mean. Our results present the expected predictions as point estimates, with uncertainty expressed through the 95% highest posterior density interval (HDI).

We pre-registered an analysis plan to assess participants' reported WTP, relative to the aggregate observed value of each prediction display. Unlike traditional Likert-style questions, WTP allows us to define an objective benchmark post-hoc for comparison to participants' responses.

Specifically, we first calculate the *expected bonus* participants gained in each condition by multiplying the expected accuracy rates by the maximum earnable bonus of $4. We then calculate the *expected bonus difference* with access to each prediction display relative to the baseline by subtracting the *expected bonus* of the baseline from the *expected bonus* of each treatment condition. We finally evaluate the discrepancy between this *expected bonus difference* and the corresponding WTPs reported by participants who used each prediction display by constructing a "*willingness-to-overpay*" metric. This measure reflects the difference between the additional bonus one is expected to earn from having access to a prediction display over the baseline and the actual difference in the average bonus participants received from accessing a prediction display relative to the baseline. We use bootstrapping to derive 95% confidence intervals with the median as a point estimate.

3.7 Qualitative Analysis of Strategies

We performed a qualitative analysis and developed a bottom-up open coding scheme using the strategies provided by the participants. Two coders worked in collaboration, each coding half of the strategies. After coding, any ambiguous strategies were discussed until mutual agreements were reached. Among 585 valid strategies, we excluded 41 uninformative ones (Code: 1) with responses such as "I have no idea, trying my hardest". For the remaining strategies, we identified 17 ways in which the prediction displays and the search features were used, as illustrated in Table 4.

Table 4: Summary of participants' strategy categories derived from open coding.

Code	Category
2–6	How do participants use different prediction displays?
7–10	Do participants find predictions to be trustworthy?
11–15	How do participants use the search feature to find a preferred label?
16–18	How do participants narrow down their choices and identify a preferred label?

We provide detailed code descriptions in Table 9 of Appendix A, and generate a spreadsheet of codes and quotes to summarize the diversity and nuances in participants' strategies, which are included in the Supplemental Material.

4 RESULTS

4.1 Data Preliminaries

We received 599 valid responses after excluding one participant from the `baseline` condition according to our pre-registered exclusion rule. As demonstrated in Figure 5 left, participants spent, on average, ≈20 minutes completing the 16 image labeling tasks. Participants in the `Top-10` and RAPS conditions took slightly longer to complete tasks than those in the `baseline` and `Top-1` conditions.
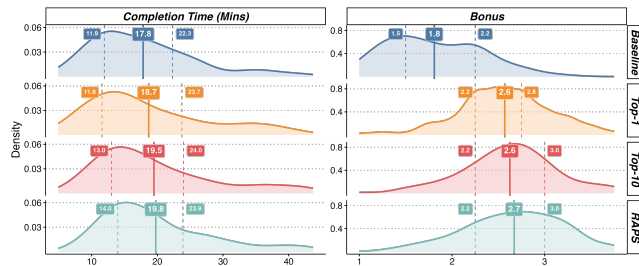

Figure 5: Distribution of participants' study completion time (minutes) and bonus earned by conditions with a solid vertical line showing average and dotted vertical lines showing the interquartile range.

From the distribution of earned bonuses shown in Figure 5 right, there is a clear difference indicating that access to the prediction displays helped participants achieve more correct answers and hence a higher bonus reward than those without access. On average, `baseline` participants provided seven correct labels, which is lower than those who used prediction displays (10 out of 16 correct).

4.2 Accuracy

4.2.1 Overview. In Figure 6, we present the expected predictions by our accuracy model (Model 1) for each type of image stimuli, with uncertainty quantified as 95% HDIs. At a high level, we find that RAPS does not improve participants' labeling accuracy for easy tasks but is more useful for hard tasks, especially for OOD images. While a `smaller` RAPS can be more useful for in-distribution images, size does not appear to affect accuracy much when images are

OOD. Because `smaller` and `larger` set sizes depend on whether instances are `in-distribution` or OOD in our data-generating model (Section 3.2), we provide the average set size for size groups to contextualize the results below.

4.2.2 In-distribution.

For both `easy` and `hard` images that are `in-distribution`, a `larger` set size may decrease participants' labeling accuracy. This effect is most prominent when images are classified as hard in Figure 6, (C) and (D). When using RAPS to label hard images, participants' accuracy is higher when the set size is `smaller` (74.2%; HDI: [66.5%, 81.3%]; average size: 2.6), and lower when set size is `larger` (52.8%; HDI: [43.5%, 60.9%]; average size: 19.5). We speculate that the reason is that a `larger` set size increases participants' cognitive load as they need to traverse through the set and evaluate all labels. Figure 6 (C) shows that expected accuracy of RAPS participants is roughly 8% higher than that of the `Top-10` condition (64.5%; HDI: [56.4%, 71.7%]) which is roughly 10% higher than the `Top-1` condition (55.6%; HDI: [47.2%, 63.7%]), though HDIs overlap.

When the true label is in the prediction display, participants do not always choose it. This is particularly evident for `in-distribution` images categorized as easy (cf. Figure 6, A and B). Here, the participants' labeling accuracy tends to be lower than the prediction display accuracy (i.e., the percentage of the time the prediction display included the true label), depicted as solid horizontal lines. However, when labeling hard images, participants in the `Top-1` condition can improve upon the model's prediction. As suggested by our open codes discussed in Section 4.5, we find that when the `Top-1` prediction was obviously inaccurate, participants were more likely to rely on their own judgment and used the search tool to identify the correct answer. The presence of a greater number of labels in the `Top-10` and RAPS conditions may have complicated the labeling process due to an increased cognitive load. This is exemplified by the observed trends in these conditions: among incorrect responses, participants in the RAPS condition missed the correct label from the prediction set 60% of the time, whereas those in the `Top-10` condition did so 80% of the time.

In summary, when task images are `in-distribution`, (1) labeling accuracy for participants in the `Top-1` or `Top-10` conditions is higher when the images are easy (cf. Figure 6, A and B); (2) the accuracy of RAPS participants is negatively correlated with set size: a `larger` set size decreases accuracy, particularly for hard images (cf. Figure 6, C and D); and (3) `Top-1` participants are more likely to rely on their own judgment and use the search tool to identify the correct answer when predictions are less likely to be correct. In contrast, the pattern of incorrect responses observed in the `Top-10` and RAPS conditions suggest that the presence of more predictions makes it harder for participants to discern between inaccurate and accurate predictions, even when the correct label is present.

4.2.3 Out-of-distribution.

When OOD images are `easy`, Top-k predictions yield higher accuracy than RAPS. Due to the similarity in HDIs across conditions for easy OOD images (cf. Figure 6, E and F), we report expected accuracy marginalized over set sizes. We find `Top-1` participants tend to achieve slightly higher accuracy in expectation (89.6%; HPI:
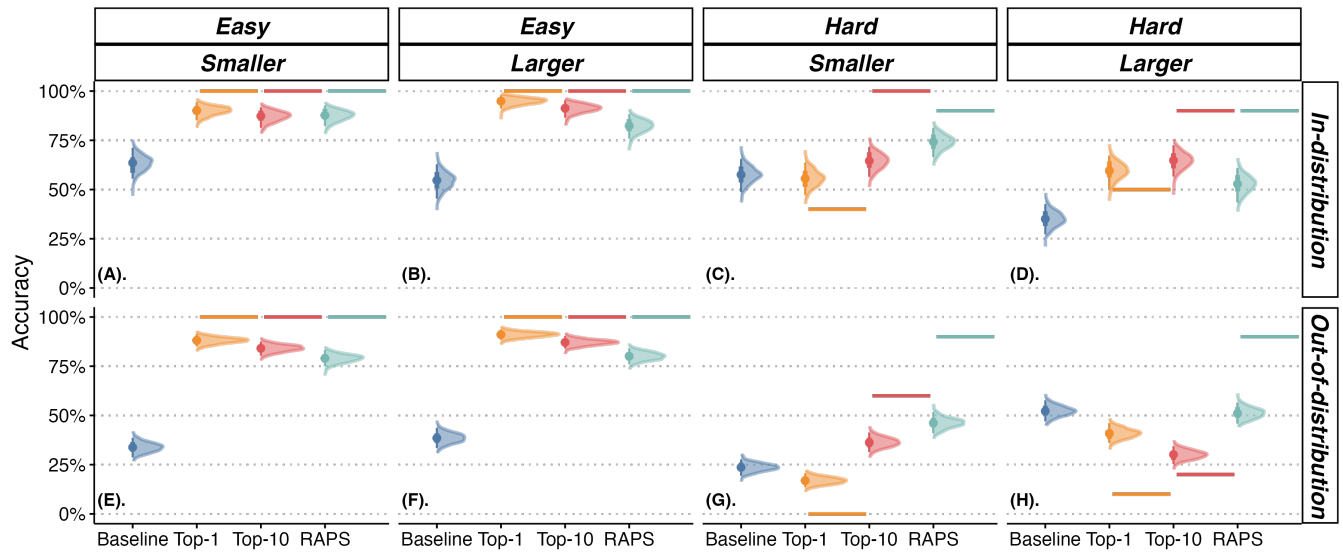
Figure 6: Median expected labeling accuracy for in-distribution and OOD task images grouped by difficulty (`easy` and `hard`) and set size (`smaller` and `larger`) predicted by the accuracy model (Model 1) with uncertainty expressed as 95% HDIs. Solid horizontal lines show the prediction accuracy achieved by that display (i.e., the probability that the displayed predictions included the true label) using our base classifier (i.e., WRN). Note: for easy in-distribution stimuli, the average set sizes are 2.6 (`smaller`) and 19.5 (`larger`); for hard, they are 8 (`smaller`) and 51.1 (`larger`). For easy OOD stimuli, the averages are 5 (`smaller`) and 28.1 (`larger`); for hard OOD, they are 30.1 (`smaller`) and 90.8 (`larger`).

[85.5%, 93.4%]) than those using `Top-10` (85.7%; HDP: [81%, 89.6%]). Meanwhile, the `RAPS` condition yields the lowest expected accuracy of 79.5% (HPI: [75.1%, 83.6%]), despite an average set size of 5 (as detailed in Table 3) that is smaller than the size of `Top-10`, with identical chance that the prediction display contained the true label (i.e., colored horizontal lines in Figure 6, E and F). This divergence from our `in-distribution` finding, where `smaller` set sizes typically correlate with higher accuracy when predictions are reliable, suggests that additional factors may be at play in the `OOD` scenario. It is possible that the reduction in cognitive load afforded by having access to softmax scores in the `Top-10` condition aided participants in finding the correct label. This observation partially reinforces our finding from the `easy in-distribution` results: when the model is well-calibrated (cf. Figure 6, E and F), viewing fewer predictions and having access to prediction confidence (i.e., as in our Top-k conditions) can improve accuracy.

When `OOD` images are `hard`, `RAPS` yield the highest labeling accuracy regardless of the set size. For hard `OOD` images, `RAPS` set sizes grow substantially larger, with `smaller` sets averaging 30 instances and `larger` sets averaging 91 instances. When hard `OOD` images had `smaller` set sizes (i.e., Figure 6, G), `RAPS` participants outperform those in Top-k conditions with a relatively high accuracy of 45.1% (HPI: [41%, 51.6%]), followed by the `Top-10` condition (36.3%; HPI: [31.4%, 41.3%]) and `Top-1` condition (16.9%; HPI: [13.5%, 20.7%]). When the set size is `larger` (i.e., Figure 6, H), we identify a consistent pattern that `RAPS` participants (51.1%, HPI: [45.9%, 56.5%]) outperform those in the Top-k conditions. Nevertheless, the accuracy of `RAPS` participants is still lower than the

average coverage rate (cf. Figure 6, G and F, with `green` solid horizontal lines), implying that even when the sets include the true label, it is still challenging for participants to identify it.

When `OOD` images are `hard`, participants relying on their own judgment may achieve better accuracy than those relying on poorly-calibrated Top-k predictions. For hard `OOD` images with `smaller` average set size (i.e., Figure 6, G), we find that participants in the `baseline` condition who did not have access to any predictions have a higher accuracy of 23.6% (HPI: [19.3%, 27.6%]) than those in the `Top-1` condition (16.9%; HPI: [13.5%, 20.7%]). For hard `OOD` images with a `larger` average set size (i.e., Figure 6, H), the accuracy of using `RAPS` (51.1%, HPI: [45.9%, 56.5%]) versus the `baseline` condition (52.2%; HPI: [47%, 57.7%]) is similar, and participants who used `Top-1` prediction are more accurate (40.9%; HPI: [36.1%, 46%]) than those who used `Top-10` predictions (30.1%; HPI: [25.3%, 34.4%]). Our results suggest that Top-k predictions from a poorly calibrated model (cf. Figure 6, G and H) can negatively influence human judgment. Although `RAPS` participants have relatively higher accuracy than those in the Top-k conditions, their accuracy is almost identical to those in `baseline`, suggesting it might be beneficial to withhold predictions for hard `OOD` stimuli that require a `larger` set size to achieve the desired coverage.

In summary, (1) when `OOD` images are easy and the predictions and softmax scores well-calibrated (cf. Figure 6, E and F), Top-k predictions tend to result in higher labeling accuracy; (2) when `OOD` images are hard and the model predictions more error-prone (cf. Figure 6, G and H), `RAPS` participants can outperform their Top-k counterparts. Unlike in the in-distribution case, increasing the size
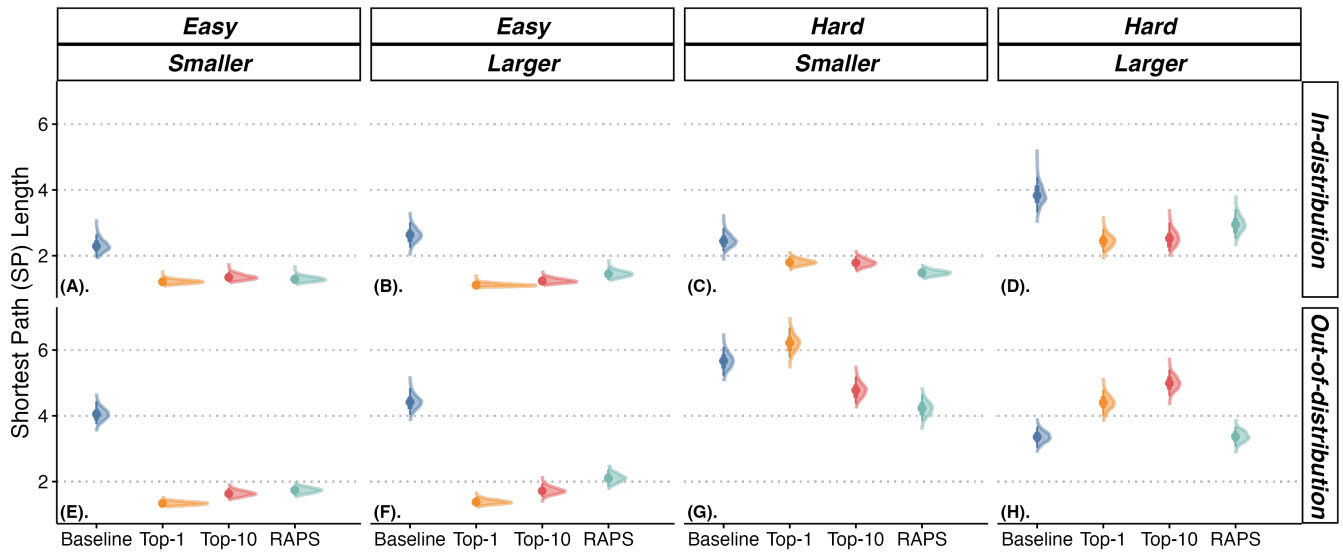
Figure 7: Median expected shortest path length quantifies the amount of label error for `in-distribution` and `OOD` task images grouped by difficulty (`easy` and `hard`) and set size (`smaller` and `larger`) predicted by the shortest path model (Model 2) with uncertainty expressed through 95% HDIs. Note: for easy in-distribution stimuli, the average set sizes are 2.6 (`smaller`) and 19.5 (`larger`); for hard, they are 8 (`smaller`) and 51.1 (`larger`). For easy OOD stimuli, the averages are 5 (`smaller`) and 28.1 (`larger`); for hard OOD, they are 30.1 (`smaller`) and 90.8 (`larger`).

of the prediction set does not degrade accuracy; and (3) when `OOD` images are `hard` and the model predictions are more error-prone, participants who rely on their own judgment and the label space search tool may be more likely to identify the correct answer. The reasons behind the performance advantage of the participants in the `baseline` condition when labeling `hard` `OOD` images with `larger` set sizes remain ambiguous. It is possible that the accuracy decrease for participants in **Top-10** condition when labeling `hard` `OOD` images that have `larger` set size is due to overreliance on inaccurate predictions, given that the average coverage for **Top-10** predictions decreases from 60% (`smaller`) to 20% (`larger`), as shown in Table 3.

4.3 Shortest Path Length

The predictions of our SP length model (Model 2) are highly correlated with those of the accuracy model when comparing the results in Figure 6 with those in Figure 7. We briefly highlight the main takeaways that complement our accuracy results, with more detailed descriptions in Appendix B.2.

When OOD images are hard, a larger prediction set leads to submitted labels that are slightly closer to the correct label. Recall that in Section 4.2.3, we find when `OOD` images are classified as `hard`, RAPS participants can outperform their Top-*k* counterparts, but there is no difference in accuracy by set size (cf. Figure 6, G and H). However, from the SP length, we find that if the task image has a `larger` set size, RAPS participants' chosen label is slightly closer to the correct answer in the label space hierarchy. As demonstrated in Figure 7 (G) and (H), the SP length for `hard` `OOD` images that have a `larger` set size is 3.4 (HPI: [3.1, 3.7], average size: 90.8) and that of a `smaller` set size is 4.2 (HPI: [3.8, 4.7]; average size: 30.1).

It is difficult to say why this might be the case; it is possible, as supported by our open codes in Section 4.5, that when `OOD` stimuli are especially challenging to recognize, participants spend more time scrutinizing labels in the prediction set for clues, and that even when the probability that the prediction set contains the true label is the same, viewing more predictions helps cue their ability to identify the true label. The effect is quite small, whatever the cause.

When predictions are generated by a well-calibrated model, having access to predictions can guide participants closer to the correct answer. This result is most prominent for `hard` `in-distribution` images with `smaller` set size. From Figure 6 (C), we see the `baseline` and **Top-1** prediction tend to yield similar accuracy, as HDIs closely overlap. However, in Figure 7 (C), we find that participants who used **Top-1** prediction can provide responses that are closer to the correct answer than those in the `baseline`.

In summary, we find (1) for `hard` `OOD` stimuli, although the accuracy of RAPS participants is similar, a `larger` prediction set may guide responses closer to the ground truth (cf. Figure 7, G and H); and (2) when the model is well-calibrated, inaccurate Top-*k* predictions might help participants in identifying the correct label.

4.4 Willingness-to-Pay

We present the distributions of elicited willingness-to-pay (WTP) measures in Figure 11 with summary statistics in Table 10 of Appendix B.3. Elicited WTPs across conditions are quite similar, suggesting that the measure may be noisy. Participants are willing to pay slightly more for **Top-10** ($1.84; CI: [1.58, 2.11]) than RAPS ($1.78; CI: [1.51, 2.07]) and **Top-1** ($1.7; CI: [1.46, 1.96]).

Nonetheless, as detailed in Section 3.6, we quantify participants' perceived utility for each prediction display using a measure we devised, called "*willingness-to-overpay*". This proxy focuses on economic valuation, offering direct monetary comparisons between the monetary value of a prediction display that participants perceive (i.e., WTP) against how much additional expected bonus participants gained from accessing each prediction display.

Table 5: *Expected Bonus Diff*. column shows the additional expected bonus participants gained when completing tasks using each prediction display relative to the `baseline`.

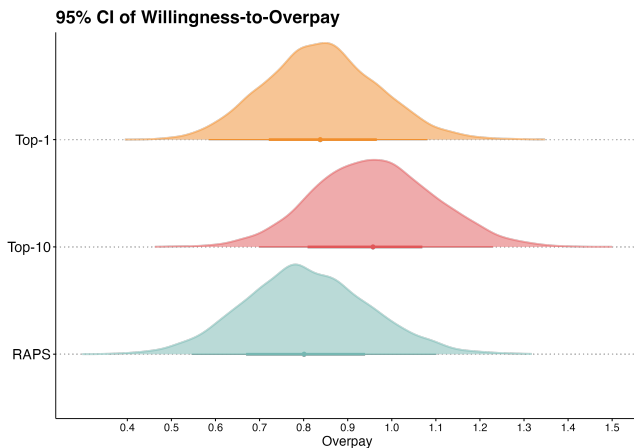| Condition	Accuracy	Expected Bonus	Expected Bonus Diff.
`baseline`	0.41	1.64	0.00
Top-1	0.63	2.52	0.86
Top-10	0.64	2.56	0.88
RAPS	0.66	2.64	0.98

Figure 8: Bootstrapped 95% CIs with the median as a point estimate for "*willingness-to-overpay*", computed by the difference between the *expected bonus differences* and the willingness-to-pay elicited from participants.

We note that this measure is imperfect as a comparator for any individual WTP because we cannot observe the counterfactual performance without a prediction display for each participant individually. In aggregate, however, Figure 8 suggests that participants may over-value the prediction displays to a similar extent across conditions. If anything, despite benefiting more from using the prediction sets, we see weak evidence that RAPS participants perceive a lower monetary value from the prediction set relative to Top-k predictions. RAPS participants are willing to overpay by $0.8 on average (CI [0.54, 1.1]), slightly less than Top-1 participants ($0.84, CI [0.598, 1.10]). Participants find Top-10 prediction most useful on average ($0.96; CI [0.7, 1.23]).

4.5 Qualitative Analysis of Strategies

From our open coding, we find that participants employed different strategies between prediction displays (i.e., Figure 10 in Appendix B.1). In general, most participants reported consulting predictions when decision-making (≈81% out of 450). Some participants reported first inspecting predictions and assessing their accuracy using their judgment (Top-1: ≈17% out of 150; Top-10: 32% out of 150; RAPS: ≈39% out of 150), while others report relying first on their own intuitions to form a general idea, and then seeing if it aligned with any predictions (Top-1: ≈23% out of 150; Top-10: 20% out of 150; RAPS: ≈23% out of 150). Very few participants chose to rely completely on their own intuitions because they find predictions inaccurate or not useful (Top-1: ≈3% out of 150; Top-10: ≈1% out of 150; RAPS: ≈1% out of 150). Others reported they would only consult predictions when the image was too hard to recognize and they had no clue (Top-1: 8% out of 150; Top-10: ≈25% out of 150; RAPS: 10% out of 150). Among participants who reported that they trusted AI prediction(s), some of them in Top-k conditions placed higher trust on predictions with high softmax pseudo-probability (e.g., ≥ 50%) (Top-1: ≈30% out of 23; Top-10: ≈54% out of 35), while those in RAPS condition focused more on the top suggestions in the prediction set (50% out of 26).

Collectively participants used all features of our interface, reporting that the dropdown (≈1% out of 599), bottom-up (≈9% out of 599), and keyword search (≈13% out of 599) methods were useful in helping them identify a label, and the representative images were informative for aiding understanding of the meaning of labels (≈14% out of 599). Of participants who viewed prediction displays, those in the Top-1 condition appeared most likely to use the label space hierarchy network to find a better match (**60% out of 150**, compared to ≈**23% out of 150** and **28% out of 150**).

In addition to our open codes, we summarized the proportion of participants who fully relied on predictions (i.e., with all submitted responses selected from predictions). Aligning with the codes described above, we find Top-1 participants relied less on the predictions and used the search tool to find their preferred choice more often for both in-distribution (≈39% out of 150) and OOD (≈3% out of 150) images. Top-10 participants reported relying heavily on predictions for in-distribution images (≈95% out of 150) and much less for OOD (25% out of 150). RAPS participants similarly reported very frequently basing their decisions on a constrained set of predictions in-distribution (90% out of 150) and slightly less but still quite often for OOD (48% out of 150).

5 DISCUSSION

Conformal prediction has gained recent visibility as a distribution-free approach for quantifying uncertainty in model predictions. Our results suggest that decision-makers can use prediction sets effectively in an AI-advised labeling task, but that their advantages over status quo Top-k presentations vary with the properties of the setting. When the model has high accuracy and the test instances are in-distribution, we find that the size of the prediction set is a critical determinant of its utility. Smaller sets lead to performance on par with or slightly better than Top-k displays while larger ones lead to slightly worse performance. In contrast, when the model encounters unforeseen and challenging OOD instances that can

cause Top-k predictions and their associated uncertainty scores to be unreliable, prediction sets offer a comparative advantage, regardless of their set size, at least in a setting like ours where coverage guarantees are maintained. We elucidate our main findings and discuss the broader implications of our study.

A smaller prediction set derived from a well-calibrated model is generally more useful. Our results suggest that the utility of conformal prediction sets is linked to set size, presumably due to an underlying factor of cognitive load. A large set that embodies more uncertainty requires users to navigate and evaluate many incorrect predictions before finding the correct answer. This increase in cognitive load generally makes prediction sets less effective and may encourage more satisficing [54], resulting in a lower response accuracy. More rigorous uncertainty quantification does not always lead to better decision-making. Designers of prediction displays should carefully consider the trade-off between set size and adaptiveness when using conformal prediction sets to communicate prediction uncertainty for machine learning models in practical settings.

Conformal prediction sets can be more useful for hard OOD instances. One possible reason prediction sets can exceed in performance for hard OOD stimuli is their adaptiveness. While the covariate shifts used in our experimental setting can severely affect the accuracy of Top-k predictions, adaptive prediction sets under certain shifts can retain coverage by significantly increasing the set size (e.g., [18]). While the high coverage of prediction sets for OOD instances does not necessarily equate to the high accuracy of AI-advised humans, as the results for hard instances in Figure 6 demonstrate, even large set sizes for OOD instances provided some value to participants. Even if participants did not select the correct label from the set, when given larger prediction sets for hard OOD stimuli, they tended to select labels closer to the correct one, as measured by the shortest path length in the label space hierarchy. Hence, our findings highlight the potential value of conformal prediction sets in enhancing the usefulness of model predictions when encountering OOD instances in real-world applications. Designers of prediction displays should leverage these sets' adaptiveness and distribution-free properties.

Withholding predictions of an uncalibrated model may improve decision quality. Consistent with prior work on AI-advised decision-making (e.g., [7, 10, 21, 38]), our results suggest that when a model is well-calibrated and more accurate than humans alone, users with access to its predictions can perform better than without the model, but not as well as the model alone for easier instances. When the model is poorly calibrated, the type of prediction display affects whether people can perform better by accessing the model predictions. For instance, with the Top-k displays showing low coverage of the true label, users sometimes performed worse than they would have done by ignoring the display. We observe qualitative feedback, such as "I used the Top-1 when it was impossible to find what the image was based on my own perception" and "In cases where I really was unsure and did not know what the picture was, I relied on the AI and went with their label". Decision-making quality, when presented with a constrained set of label options, is highly correlated with the calibration of prediction information.

Only those using RAPS matched the expected performance of having no prediction display in hard OOD instances with larger set sizes. Our results underscore the potential value of treating the detection of instance type as a model's learning problem (e.g., [32]).

5.1 Limitations and Future Work

It is unclear whether a larger prediction set showing more possible labels can still be effective if the coverage is reduced. A natural follow-up study is to evaluate the efficacy of prediction sets in an OOD setting where coverage is disrupted at varying magnitudes. The adaptiveness of the prediction sets [3, 52] enhances a model's resilience to random and common perturbations that can occur in real-world scenarios, such as image corruptions. However, future work may explore the use of intentional adversarial methods that can attack softmax (e.g., FGSM [20] or PGD [39]). Combined with the finding from Straitouri and Rodriguez [56] that constrained choices can lead to better performance based on sets derived from a calibrated model with a smaller size, we might see an opposite effect to what we observed, where people are less likely to pick a choice from a prediction set that has a lower chance of containing the true label. While our experiment isolated the effect of the conformal prediction set and guarantee without performing other interpretability manipulations to the Top-k conditions, future work might also consider comparing conformal prediction sets to post-hoc calibrated softmax with Top-k displays.

Our experimental design, which does not observe individual participant performance without a prediction display, makes willingness-to-pay an imperfect comparator. Future work should delve into how users perceive the value of prediction displays and conduct within-subject comparisons of both performance and perceived value. Additionally, it is unclear whether the conformal coverage guarantee is prone to misconceptions common to conventional confidence intervals, such as misinterpreting the coverage rate and disregarding labels not included within the set.

6 CONCLUSION

Uncertainty quantification for deep neural networks (NNs) has been challenging due to their black-box nature. We evaluate whether communicating prediction uncertainty via conformal prediction sets can improve human decision-making in AI-advised image labeling. We contribute the results of a large online experiment, in which we evaluate the utility of conformal prediction sets against status quo Top-k predictions across a diverse range of stimuli, varied by in-distribution or OOD, level of difficulty, and set size. Our work shows that when the model is well-calibrated, a smaller prediction set is most beneficial, as larger sets can overwhelm users and decrease performance due to increased cognitive load. However, when the model faces unexpected OOD instances, larger prediction sets can be more useful than their Top-k counterparts, at least when coverage remains high. Our work sheds light on how conformal prediction sets can be used to rigorously quantify uncertainty for machine learning models in practical applications and outlines potential avenues for future research.

REFERENCES

[1] Sandro Ambuehl and Shengwu Li. 2018. Belief updating and the demand for information. *Games and Economic Behavior* 109 (2018), 21–39.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193* (2020).

[4] Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).

[5] Varun Babbar, Umang Bhatt, and Adrian Weller. 2022. On the utility of prediction sets in human-AI teams. In *Proceedings of the Thirty-First International Joint Conferenceon Artificial Intelligence (IJCAI-22)*. 2457–2463.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 1–16.

[7] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 1–12.

[8] Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the driver–automation interaction: An approach using automation uncertainty. *Human Factors* 55, 6 (2013), 1130–1141.

[9] Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32, 1 (2006), 13–47.

[10] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. AAAI Press, 95–106.

[11] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological Science* 25, 1 (2014), 7–29.

[12] Geoff Cumming and Sue Finch. 2005. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist* 60, 2 (2005), 170–180.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[14] Rajmadhan Ekambaram, Dmitry B Goldgof, and Lawrence O Hall. 2017. Finding label noise examples in large scale datasets. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2420–2424.

[15] Christiane Fellbaum. 1998. *WordNet: An electronic lexical database.* MIT Press.

[16] Hironobu Fujiyoshi, Tsubasa Hirakawa, and Takayoshi Yamashita. 2019. Deep learning-based image recognition for autonomous driving. *IATSS research* 43, 4 (2019), 244–252.

[17] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. Bayesian workflow. *arXiv preprint arXiv:2011.01808* (2020).

[18] Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. 2022. Adversarially robust conformal prediction. In *Proceedings of the ICLR*.

[19] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. 2023. Conformal Prediction With Conditional Guarantees. *arXiv preprint arXiv:2305.12616* (2023).

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *Proceedings of the ICLR*.

[21] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *In Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, International Conference on Machine Learning (ICML), 1321–1330.

[23] Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. 2024. A Statistical Framework for Measuring AI Reliance. *arXiv preprint arXiv:2401.15356* (2024).

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[25] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[26] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *In Proceedings of the ICLR*.

[27] Rohan Hore and Rina Foygel Barber. 2023. Conformal prediction with local weights: randomization enables local guarantees. *arXiv preprint arXiv:2310.07850* (2023).

[28] Jessica Hullman, Alex Kale, and Jason Hartline. 2024. Decision Theoretic Foundations for Experiments Evaluating Human Decisions. *arXiv preprint arXiv:2401.15106* (2024).

[29] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 903–913.

[30] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS One* 10, 11 (2015), e0142444.

[31] Susan L Joslyn and Jared E LeClerc. 2012. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied* 18, 1 (2012), 126–140.

[32] Ramneet Kaur, Kaustubh Sridhar, Sangdon Park, Yahan Yang, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. 2023. CODiT: Conformal out-of-distribution Detection in time-series data for cyber-physical systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023) (ICCPS '23)*. ACM, 120–131.

[33] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 ACM CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.

[34] Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. *Master's Thesis, University of Toronto* (2009).

[35] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems*. ACM, Article 54, 1–18 pages.

[36] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. ACM, 29–38.

[37] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *In Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[40] Charles F Manski. 2019. Communicating uncertainty in policy analysis. *In Proceedings of the National Academy of Sciences* 116, 16 (2019), 7634–7641.

[41] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484* (2019).

[42] George A Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[43] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *Proceedings of the CVPR workshops*, Vol. 2. IEEE.

[44] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[45] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *Proceedings of the Machine Learning: ECML 2002: 13th European Conference on Machine Learning*. Springer Berlin, Heidelberg, 345–356.

[46] Adam Paszke et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019).

[47] John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 10, 3 (1999), 61–74.

[48] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, 379–396.

[49] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning.* Mit Press.

[50] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. 2018. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making* (2018), 323–350.

[51] Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1 (IJCAI'95)*. Morgan Kaufmann Publishers Inc., 448–453.

[52] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems* 33 (2020), 3581–3591.

[53] Christoph Sager, Christian Janiesch, and Patrick Zschech. 2021. A survey of image labelling for computer vision applications. *Journal of Business Analytics* 4, 2 (2021), 91–110.

[54] Herbert A Simon. 1956. Rational choice and the structure of the environment. *Psychological Review* 63, 2 (1956), 129–138.

[55] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *In Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

[56] Eleni Straitouri and Manuel Gomez Rodriguez. 2023. Designing decision support systems using counterfactual prediction sets. *arXiv preprint arXiv:2306.03928* (2023).

[57] Barry N Taylor and Chris E Kuyatt. 1994. *Guidelines for evaluating and expressing the uncertainty of NIST measurement results.* Vol. 1297. National Institute of Standards and Technology.

[58] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the CVPR.* IEEE, 1521–1528.

[59] Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin* 76, 2 (1971), 105.

[60] Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. 2021. Biological data annotation via a human-augmenting AI-based labeling system. *NPJ Digital Medicine* 4, 1 (2021), 145.

[61] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world.* Vol. 29. Springer.

[62] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference 2016.* British Machine Vision Association, BMVA Press.

[63] Dongping Zhang, Eytan Adar, and Jessica Hullman. 2021. Visualizing uncertainty in probabilistic graphs with Network Hypothetical Outcome Plots (NetHOPs). *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 443–453.

[64] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency.* ACM, 295–305.

[65] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. 2017. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *Proceedings of the 16th IFIP TC 13 International Conference on Human-Computer Interaction — INTERACT 2017.* Springer-Verlag, 23–39.

A FURTHER METHODOLOGICAL DETAILS

We present the distribution of the cross-entropy loss for stimuli that are in-distribution in Figure 9. Table 6 shows WRN's Top-k prediction accuracy and conformal coverage for in-distribution stimuli and 15 types of covariate shifts. Table 8 outlines the sampling correction for images with a deterministic set, while Table 9 explains the open codes used to categorize participants' strategies. Furthermore, Table 7 compares the performance of the Human-AI team, humans working alone, and AI standalone accuracy.
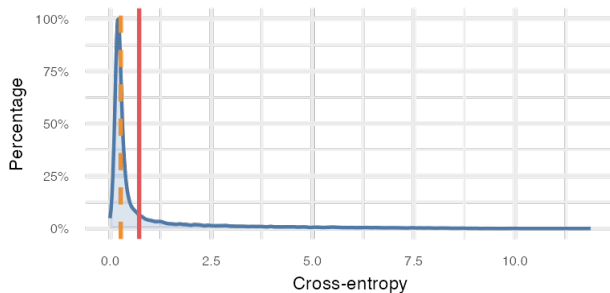

Figure 9: The distribution of cross-entropy loss observed across in-distribution predictions. The orange dotted line indicates the median when including images with a set size of one. In contrast, the red solid line indicates the median after excluding these images.

Table 6: We evaluated WRN's Top-k prediction accuracy and conformal coverage across 15 types of covariate shifts as discussed by Michaelis et al. [41]. For our experimental design, we specifically chose the corruption types marked in bold that reduce WRN's Top-1 prediction accuracy the most as OOD stimuli. The exception is the "none" category, which we used as in-distribution stimuli.

Corruption	Coverage		
	Top-1	Top-10	RAPS
none	0.82	0.98	0.96
brightness	0.77	0.96	0.95
contrast	0.76	0.96	0.95
fog	0.72	0.93	0.94
jpeg compression	0.7	0.93	0.94
pixelate	0.66	0.9	0.92
motion blur	0.57	0.83	0.88
impulse noise	0.57	0.84	0.89
shot noise	0.56	0.83	0.89
gaussian noise	0.55	0.83	0.89
elastic transform	0.52	0.78	0.84
defocus blur	0.52	0.81	0.91
snow	0.51	0.79	0.86
frost	0.43	0.7	0.76
zoom blur	0.38	0.66	0.73
glass blur	0.29	0.55	0.72

Table 7: Accuracy achieved by Human-AI teams, humans alone, and AI alone when stimuli were varied by in- or out-of-distribution (i.e., *covariate shift*).

Condition	Covariate Shift	Human+AI	AI	Human
Top-1	in	0.74	0.73	0.53
	out	0.59	0.52	0.38
Top-10	in	0.76	0.97	0.53
	out	0.59	0.7	0.38
RAPS	in	0.73	0.94	0.53
	out	0.63	0.95	0.38

B ADDITIONAL RESULTS

B.1 Qualitative Analysis of Strategies

Figure 10 presents a trellis plot summarizing the qualitative results per identified open code. Each row presents a code category with columns showing a barchart summarizing counts of each unique code from participants' strategies by treatment conditions.

B.2 Shortest Path Length

The *shortest path (SP) length* quantifies the "incorrectness" of participants' labeling choices relative to the correct label in the label space hierarchy network. Similar to accuracy, we analyze the variations of participants' labeling errors by task image types. We report the median of expected predictions for each type of image stimuli with

Table 8: The table demonstrates the sampling correction performed to include images with deterministic set for groups having `smaller` set size. For each group varied by *Corruption* and *Difficulty,* we compute a *Size Ratio,* which captures the ratio of images with deterministic set (*Size 1 Count*) to uncertain set (*SizeN Count*). *SizeN Keep* shows the number of images with uncertain set whose sizes fall near the median size in that group (i.e., 45th and 55th percentiles). We finally append images with deterministic set (*Size1 Included*) back to the group proportionally according to *Size Ratio*.

Corruption	Difficulty	Size1 Count	SizeN Count	Size Ratio	SizeN Keep	Size1 Included
none	hard	539	2794	0.19	749	**144**
none	easy	13428	2962	4.53	1778	**8060**
defocus blur	hard	331	7750	0.04	870	**37**
defocus blur	easy	4571	2431	1.88	385	**724**
snow	hard	664	7319	0.09	827	**75**
snow	easy	4964	2554	1.94	451	**877**
frost	hard	993	9073	0.11	996	**109**
frost	easy	2896	1590	1.82	249	**454**
zoom blur	hard	722	9013	0.08	1066	**85**
zoom blur	easy	2925	1914	1.53	235	**359**
glass blur	hard	516	9973	0.05	1108	**57**
glass blur	easy	2025	1360	1.49	331	**493**

Table 9: Open codes used to summarize participants' strategies.

| Category	Open Codes	Description
1	UNINFORMATIVE	Content that doesn't provide any significant or relevant information
2	HUMAN-AI COMPARISON	Participant analyzes differences and similarities between own's decisions and AI predictions
3	HUMAN KNOWLEDGE	Participant makes decisions based solely on own intuition, wisdom, or experience
4	HUMAN KNOWLEDGE 1ST & AI ADVICE 2ND	Participant makes decision based on own's intuition and experience first, supplemented by AI advice or suggestions
5	AI ADVICE 1ST & HUMAN CONFIRM 2ND	Participant seeks advice from AI first and then confirms with own's knowledge
6	AI ADVICE WHEN HARD	Participant seeks advice from AI only and specifically in challenging or complex scenarios (e.g., blurry images, etc.)
7	TRUST AI PREDICTION(S)	Participant has confidence in the prediction made by AI
8	TRUST AI (RANKING OR CONFIDENCE)	Participant relies on AI's ranked suggestions or its confidence level in predictions. Participant mentions looking only at top predictions in RAPS and Top-10, while in Top-1 referring to specific confidence percentages
9	TOP PREDICTION FALLACY	Participant expects to see the true answer in top predictions, but they believe the top prediction is wrong
10	DISAPPOINTED AT AI	Participant specifically mentions dissatisfaction with AI's performance and does not trust AI's prediction
11	SEARCH	Participant performs a general search without specifying
12	SEARCH (BY DROPDOWN)	Participant types keywords in the input field and refers to the dropdown for matches
13	SEARCH (BY KEYWORD)	Participant clicks "Search" to view a hierarchy tree that highlights matching categories
14	SEARCH (BY AI LABEL)	Participant clicks an AI-predicted label for a bottom-up search
15	SEARCH (HIERARCHY NETWORK)	Participant expands or subtracts nodes/categories and navigates through the provided hierarchy network
16	ELIMINATION	Participant tries to narrow down their options (to search in) by elimination
17	REPRESENTATIVE IMAGES	Participant compares the task image against the given label-representative images
18	COLOR & SURROUNDINGS	Participant relies on color, shapes, sizes, and surrounding objects to figure out the correct label
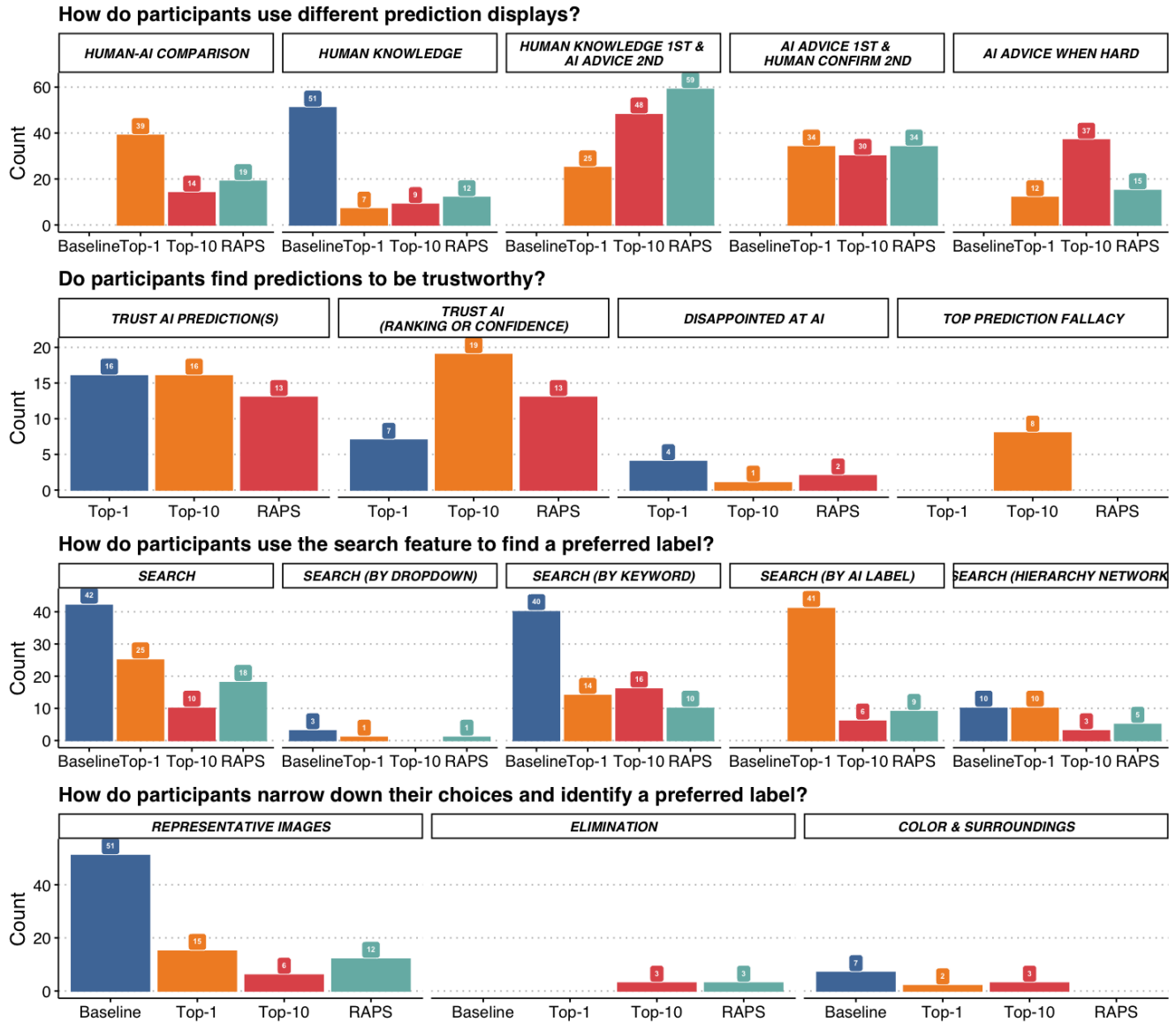
Figure 10: Summary of the qualitative results per identified open code.

uncertainty quantified as 95% HDI, holding the trial effect constant at the mean.

B.2.1 In-distribution. At a high level, the SP length is, on average, smaller for easy images across conditions. When task images are hard, the SP length is greater for images with larger set sizes across conditions.

When task images are easy, the SP length does not appear to be affected by set size—the distributions of expected SP length closely overlay across conditions (cf. Figure 7, A and B). After marginalizing by set sizes, we find that when participants have access to predictions, even when the provided label is wrong, the incorrect

label appears to be closer to the correct label in the distance (HPI: [1.05, 1.6]) than those of the `baseline` condition (HPI: [2.01, 2.93]).

When task images are hard but with a smaller average set size of 8, Figure 7 (C) shows the SP length is smallest for `RAPS` condition with a median of 1.48 (HPI: [1.34, 1.64]). We do not detect differences between the two Top-k conditions, as HPIs closely overlap (`Top-1`: 1.8 HPI: [1.63, 1.99] and `Top-10`: 1.78 HPI: [1.57, 2]). However, when without predictions, the `baseline` participants tend to submit more incorrect answers, reflected by the largest SP length of 2.44 (HPI: [2.13, 2.84]).

For hard task images with a larger average set size of 51, all intervals in Figure 7 (D) get wider than those in (C), indicating greater variations. Although SP length increased across conditions,
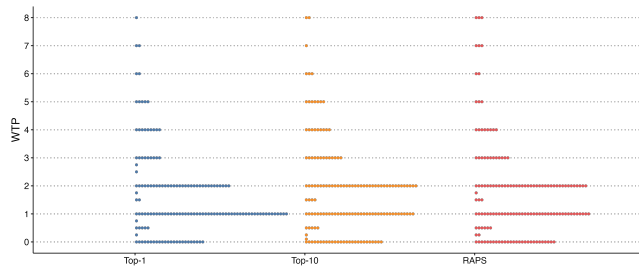
Figure 11: Distribution of willingness-to-pay as elicited from participants.

we do not detect noticeable differences among conditions with predictions (HPI: [2.98, 3.42]). Similarly, the `baseline` participants produce more inaccurate responses with a median SP length of 3.83 (HPI: [3.31, 4.39]).

In summary, we find that the SP length mirrors the accuracy in that prediction display that yields higher accuracy, which tends to produce a lower SP length. Between accuracy and SP length, a noticeable difference can be detected from task images that are hard with a `smaller` set size: although participants in the `Top-1` and the `baseline` conditions have similar labeling accuracy, the SP length of `Top-1` participants is noticeably lower than that of the `baseline`. This implies that predictions can help elicit responses closer to the correct label, even if they are wrong. Otherwise, our SP results are mostly consistent with our accuracy results.

B.2.2 Out-of-distribution. For `OOD` task images, the SP length between easy and hard images becomes more contrasting. For easy images, because predictions are still reliable and accurate, we find that the `Top-1` condition tends to produce the lowest SP length of 1.35 (HPI: [1.24, 1.49]) than `Top-10` (1.67; HPI: [1.5, 1.9]) and `RAPS` (1.89; HPI: [1.6, 2.32]), after marginalizing over `smaller` and `larger` set sizes (cf. Figure 7, E and F).

For hard `OOD` task images, we see an apparent mirroring effect relative to accuracy. When image stimuli have a `smaller` average set size of 30 shown in Figure 7 (G), `Top-1` participants have the lowest accuracy rate and also provide labels that have the highest SP length of 6.22 (HPI: [5.77, 6.68]). `RAPS` participants have the highest accuracy for this type of image and produce the lowest SP length of 4.23 (HPI: [3.84, 4.65]). The `Top-10` condition overlaps with `RAPS` with an SP length of 4.78 (HPI: [4.37, 5.19]), and both `Top-10` and `RAPS` produce lower SP lengths than `Top-1` and the `baseline` condition (5.67; HPI: [5.22, 6.1]).

Similarly, for hard `OOD` task images with a `larger` average set size of 91 shown in Figure 7 (H), `baseline` and `RAPS` conditions that can help participants achieve the highest labeling accuracy, also support them to give answers closer to the correct label with a median SP length that ranges between 3.03 and 3.68, followed by `Top-1` (4.4; HPI: [3.99, 4.79]) and `Top-10` conditions (4.99; HPI: [4.61, 5.41]).

However, despite that participants of the prediction set have similar accuracy for hard `OOD` images with `smaller` and `larger` sets, we find SP length is smaller for images with `larger` set sizes. This implies that a `larger` set size is helpful for hard `OOD` images, guiding participants to labels closer to the ground truth.

B.3 Distribution of Willingness-to-Pay

Table 10 presents the summary statistics of the elicited willingness-to-pay, while Figure 11 visualizes the distributions using a dot plot.

Table 10: Summary statistics of elicited willingness-to-pay.

Condition	Min.	Q1	Q2	Q3	Max.	Mean	SD
RAPS	0	1	1.0	2	8	1.78	1.74
Top-1	0	1	1.0	2	8	1.70	1.56
Top-10	0	1	1.5	2	8	1.84	1.67