Proceedings of the Twelfth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024)

Combining Human and AI Strengths in Object Counting under Information Asymmetry

Songyu Liu, Mark Steyvers

University of California, Irvine
songyul4@uci.edu, mark.steyvers@uci.edu

Abstract

With the recent development of artificial intelligence (AI), hybrid human-AI teams have gained more attention and have been employed to solve all kinds of problems. However, existing research tends to focus on the setting where the same task is given to humans and AI. This work investigates a scenario where different agents have access to different types of information regarding the same underlying problem. We propose a probabilistic framework that combines the predictions of humans and AI based on the quality of the information given by both agents. We apply this framework to a regression task in which humans and AI are given different views of a jar and aim to estimate the number of objects in it. We demonstrate that our model can outperform methods that ignore information asymmetry. Furthermore, we show that complementarity can be achieved, i.e., combining human and AI predictions leads to better performance than relying on humans or AI alone. This framework can be adapted to solve other problems in which different sources of information from multiple agents are present.

Introduction

As AI becomes more prevalent in computer vision and natural language processing (Recht et al. 2019; Ribeiro et al. 2020), there is a growing interest in combining human and AI capabilities in various contexts, such as crowdsourcing. A task can be assigned to a combination of human and AI workers (Kobayashi, Wakabayashi, and Morishima 2021), and an AI agent can solve consensus tasks more efficiently by learning about the available workers (Kamar, Hacker, and Horvitz 2012). Despite the recent advances in AI, there is plenty of evidence that AI can make certain types of errors that humans could easily avoid (Serre 2019). When AI fails in image classification tasks or shows demographic bias, crowdsourcing can help address these issues (Zhang et al. 2021; Zong et al. 2023). This paper explores a scenario of *information asymmetry* where a human and an AI work on the same task but have access to different types of information. As a real-world example, in the context of autonomous vehicles, the human has a view of the surroundings, while the AI driving assistant may have sensors that detect what the human cannot see (Hemmer et al. 2024).

To explore information asymmetry in a more controlled environment, our investigation focuses on the classic problem of estimating the number of objects in a jar (Surowiecki 2004). In our version of this problem, a human and an AI look at the same jar from either the same angle or two different angles. Some example images are shown in Figure 1. The jars can contain cylinders, disks, or spheres; each type of object can be small (S) or large (L). For example, cylinder (S) refers to small cylinders. Each jar can be viewed from five different angles, ranging from 0° to 90°. After looking at the jar, the AI receives additional information about the human's viewing angle and estimate. It can use that information, combined with knowledge of its own viewing angle, to determine the relative quality of the human and AI estimates. The AI's final task is to combine the human estimate with its estimate to arrive at the most accurate answer. Figure 2 shows an overview of the setup.

In our setup, the AI combines the estimates, but we can first think about how a human would combine estimates from the viewpoints shown in Figure 2 and then apply those ideas to our AI. For humans, the top view directly shows most of the individual cylinders, making estimation easier, whereas a side view makes estimation more difficult. When combining the two estimates, humans would put more weight on the estimate from the top view. More generally, a certain view may be advantageous for certain tasks but not others. A jar containing disks or spheres, for example, would appear more or less the same from the top, whether it was half full or nearly full, even though the same perspective is advantageous for counting cylinders. Our AI captures these intuitions by assessing the relative quality of the estimates using a probabilistic framework. In this framework, the probability distributions of the human and AI estimates are conditioned on the ground truth, which may be latent. Previous research (Trick, Rothkopf, and Jäkel 2023) utilized a similar model to combine experts' probability estimates. However, there are significant differences: in (Trick, Rothkopf, and Jäkel 2023), the ground truth was binary, and the mean and variance of the distributions depended solely on the individual. In our framework, the ground truth can be any positive integer, while the mean and variance are determined by the agents' perspectives.

There has been a lot of research on how AI can incorporate information from humans to solve various tasks, such

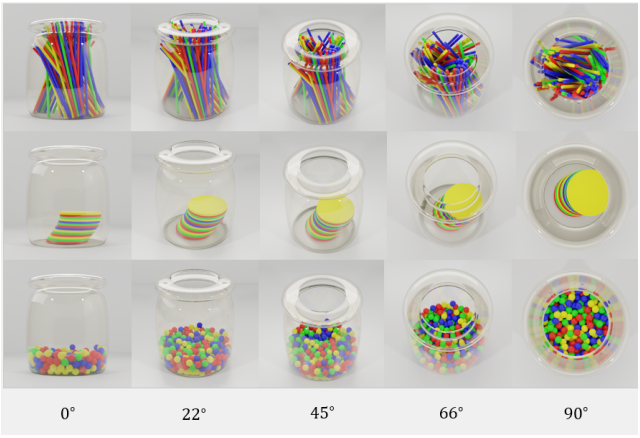

Figure 1: Examples of jar images, filled with cylinders, disks, and spheres. Each row has the same shape and each column is viewed from the same angle. The five viewing angles are listed below the images. All objects in this figure have size S.

as classification (Steyvers et al. 2022; Yang et al. 2024). However, prior work tends to focus on situations where the same information is given to humans and AI, and information asymmetry is rarely explored. Of particular interest is a study that conducted synthetic experiments in which humans and AI had access to common and unique features in a linear regression problem (Rastogi et al. 2023). All the features were independent and identically distributed (i.i.d.), which is rarely true in real-world scenarios. Our research focuses on a concrete image regression problem where agents implicitly extract features from images to make predictions, and these features are unlikely to be i.i.d.

Our main contributions are as follows:

- We propose a probabilistic framework combining asymmetric information from humans and AI to achieve better quantity estimation. The framework can determine the quality of information from various sources based on the object in the jar and the viewing angle.
- We evaluate our framework on a regression task in which humans and AI share the same goal of counting the number of objects in a jar but may have access to different views of it. Our framework outperforms baselines that do not account for information asymmetry and achieves complementarity: combined predictions are more accurate than predictions made by any single agent (Steyvers et al. 2022; Rastogi et al. 2023).

Our code and data are available at https://github.com/songyul2/AAAI-HCOMP.

Related Work

Aggregating Human Judgments

Crowdsourcing is often used to create labeled data with ground truth for supervised learning. This has led to extensive research on how to effectively aggregate human judgments (Vaughan 2018). Expert and non-expert knowledge

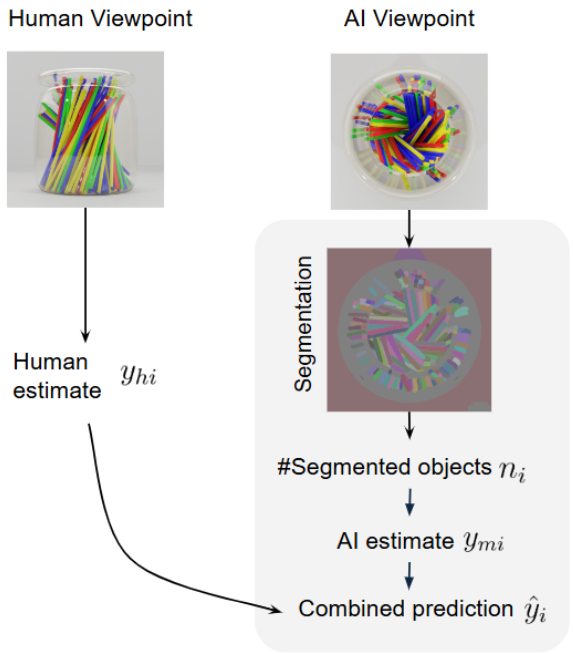

Figure 2: Overview of the pipeline. A human and an AI are given two images of the same jar taken from either the same angle or from two different angles and each individually predicts the number of objects in it. The AI then takes the human estimate into account and combines both estimates to give a final answer.

can be combined to estimate the truth using methods like Bayesian cultural consensus theory (Merkle and Steyvers 2011; Anders, Oravecz, and Batchelder 2014; Mayer and Heck 2023) and accuracy-weighted aggregation (Kameda, Toyokawa, and Tindale 2022). In the context of forecasting, weighting schemes have also incorporated factors such as forecaster bias, variance, and predictability of the forecasted quantity (Merkle, Saw, and Davis-Stober 2020). In addition, aggregation has been studied in a multimodal setting, where information is gathered from various sources (Lahat, Adali, and Jutten 2015) including continuous values and rankings from humans(Kemmer et al. 2020; Escobedo, Moreno-Centeno, and Yasmin 2022; Yoo et al. 2024). The advantage is that each modality provides additional information that cannot be obtained through others (Lahat, Adali, and Jutten 2015). We draw inspiration from these works in the multimodal setting but focus on a scenario involving different types of agents, namely humans and AI.

Human-AI Collaboration

Works in this area can be categorized based on the complementing flow: human complementing AI, AI complementing human, and bidirectional complementing (Zahedi and Kambhampati 2021). In the former two cases, one of the agents, either the AI or the human, takes on the responsibility of carrying out the task, while the other takes on the role

of an advisor. In the bidirectional setting, the agents may take turns carrying out tasks. This last case is beyond the scope of this paper and we will review works in the former two scenarios.

Our work is in the first case. Within this specific setting, prior work studied how to combine probabilities of whether certain events would occur (Benjamin et al. 2023). Another work considered how humans and AI could collaborate in optimization tasks (Tan, Gupta, and Xu 2022). In these works, all agents contribute simultaneously to the same task. Alternatively, there is also a line of work on delegating tasks (Lubars and Tan 2019; Pinski, Adam, and Benlian 2023). For example, one study proposed a greedy algorithm that let a machine learning model outsource part of a data set to humans to achieve better team performance (De et al. 2020).

In the context of AI complementing human, researchers investigated complementary expertise in controlled studies, where the AI system excels in tasks that the human partner tends to perform less accurately (Zhang, Lee, and Carter 2022). Participants adjusted their reliance on AI according to its expertise, leading to improved team performance. In our task, complementary expertise arises naturally. Later on, our results will show that humans and AI have different strengths and weaknesses when viewing various types of objects in the jar. To combine estimates, our AI learns the appropriate level of reliance on the human estimate. There is a closely related study where participants made an initial guess and then adjusted their prediction after receiving the AI judgment (Hemmer et al. 2024). Two experiments were conducted under this setup. The first one studied information asymmetry in a house price prediction task. Researchers provided tabular data to an AI and the same data with a house image to a subset of the participants. The second one constructed an AI that often provided accurate predictions for images that posed challenges for humans, resulting in the complementary expertise described above. This work and ours both address information asymmetry but our study is centered around AI and modeling, whereas (Hemmer et al. 2024) focuses on conceptualization and empirical insights.

Combining Estimates

Data-generating Process

We first introduce some terminology. We use the term "agent" to refer to either a human or an AI and "shape" to denote the type of object in the jar. Our model is designed to work with any number of agents, but we focus on the scenario involving one human and one AI. Additionally, the data-generating process remains the same for each agent and each shape, so we will focus on how a human estimate is generated given a specific shape. Suppose for image i, the ground truth is y_i number of objects, and the human has a viewing angle $a_{hi} \in [0, 90]$ and gives an estimate y_{hi}. In the model, the ground truth is used to generate an estimate according to a normal distribution with some parameters. These parameters affect either the mean or the variance of the distribution. We assume that all human individuals behave similarly, so they share the same parameters, but this assumption can be dropped without significantly changing our model. Due to the distinct characteristics of different shapes, we use a separate set of parameters for each shape.

As discussed earlier, for humans, counting cylinders is likely to be easier from the top than from the side, and the angle is a key factor. Our model captures this intuition and makes it explicit that the quality of the estimates depends on the angle. Previous research has shown that humans systematically underestimate numbers of objects (Taves 1941). Therefore, we assume that human estimates have systematic bias $b(a_{hi}; b_h)$ that depends on the viewing angle and is parameterized by b_h. In addition, a human may produce a different estimate if they see the same image a second time. This is captured by variance, which describes the amount of uncertainty associated with an estimate. Since disadvantageous views may lead to more uncertainty, the variance is a function of the viewing angle parameterized by v_h, and we denote it by $v(a_{hi}; v_h)$. Thus, we have the data-generating process:

$$y_{hi} \mid y_i, b_h, v_h \sim \mathcal{N}(b(a_{hi}; b_h) + y_i, v(a_{hi}; v_h)) \quad (1)$$

For the AI, the normal distribution that generates their estimate is similar. We have the AI estimate y_{mi}, their viewing angle $a_{mi} \in [0, 90]$, and parameters b_m and v_m. For brevity, we omit the notation of the exact distribution.

Model Inference

We split our data into training and test data sets. Details on data splitting will be explained later. The ground truth y_i is available in the training stage but not the testing stage. For both stages, estimates and angles are always available. We use the L-BFGS algorithm provided in Stan on the training data set to obtain the posterior mode of the parameters (Stan Development Team 2024). During the testing stage, we fix the parameters and infer the posterior distribution of y_i conditioning on parameters, estimates, and angles.

When we evaluate the performance of an agent by itself, one truth y_i generates one estimate y_{hi} or y_{mi} depending on whether the human or the AI is present. For humans, we have the following closed-form solution:

$$y_i \mid y_{hi}, b_h, v_h \sim \mathcal{N}(y_{hi} - b(a_{hi}; b_h), v(a_{hi}; v_h)) \quad (2)$$

A similar formula for the AI can be obtained by replacing those variables associated with humans. When both agents' estimates are given, a truth y_i generates two estimates y_{hi} and y_{mi} according to the learned parameters. This also admits a closed-form solution. We first simplify the notations. Let $y_1 = y_{hi} - b(a_{hi}; b_h), v_1 = v(a_{hi}; v_h), y_2 = y_{mi} - b(a_{mi}; b_m), v_2 = v(a_{mi}; v_m)$. Essentially, these terms are the posterior mean and variance of the truth when agents independently produce estimates. The posterior on y_i given estimates from both agents is

$$y_i \mid y_{hi}, b_h, v_h, y_{mi}, b_m, v_m \sim \mathcal{N}\left(\frac{y_1 v_2 + y_2 v_1}{v_1 + v_2}, \frac{v_1 v_2}{v_1 + v_2}\right) \quad (3)$$

The derivations can be found in Supplemental materials. If we do not assume normal distributions for the data-generating process, the posterior distributions above may not

be in closed form. However, we can still use Stan or other software to estimate the posterior distributions. From this formula, we see that the variances of the two agents are directly related to the weight given to their estimate. When combining the estimates, the AI leans towards the one given by the agent with less variance. In all cases, the posterior of the truth is a normal distribution, and we take the posterior mode as the prediction. Depending on whose estimate or estimates are given, we obtain the human prediction, the AI prediction, or the combined prediction provided by the AI after incorporating the human's judgment.

Methods

Bias and Variance as Functions of Angles

Recall that bias and variance are both functions of the angle. In the data set used for our experiments, there are five angles. Intuitively, our model needs to learn each function using these five data points. To avoid overfitting, we first restrict the family of functions to be polynomial functions. Polynomials are characterized by their degree, and we want to choose a degree that can balance complexity and flexibility. If we choose degree 4 polynomials, there will be five coefficients for five data points. If we use degree 3 polynomials with four coefficients, the complexity of the polynomials will be reduced, and it will be less likely to overfit. Thus, we have the following functions:

$$b(a_{hi}; b_h) = b_{h1} + b_{h2}a_{hi} + b_{h3}a_{hi}^2 + b_{h4}a_{hi}^3 \quad (4)$$

$$v(a_{hi}; v_h) = v_{h1} + v_{h2}a_{hi} + v_{h3}a_{hi}^2 + v_{h4}a_{hi}^3 > 0 \quad (5)$$

Obtaining AI Estimates

In this section, we explain how to obtain AI estimates from images shown in Figure 2. A natural way to estimate the number of objects in a jar is to start by counting how many objects are visible in the image. This was once a difficult problem that could not be easily accomplished by computer vision models alone (Russakovsky, Li, and Fei-Fei 2015). The recent Segment Anything Model (SAM) has made it possible and gained a lot of attention in computer vision (Kirillov et al. 2023). It can detect objects in images without human intervention. We choose HQ-SAM, which builds on SAM with superior zero-shot segmentation performance: the ability to segment objects without prior training on new categories (Ke et al. 2023). It identifies meaningful parts of an image and an example is in Figure 2, but sometimes these are not the object we want to count. In Figure 2 we can see that reflections on the jar are also recognized. This issue will be further addressed below. We also experimented with an alternative method that counts objects in images using density maps (Liu et al. 2022). However, the maps are not intuitive enough to effectively visualize what is being counted. Some segmentation models also support language prompts such as the one in (Zou et al. 2023), but we stick to those without prompting.

An important note here is that segmentation models are not designed to estimate the number of objects in the jar. The number of segmented objects only approximates the number of visible objects. To bridge this gap, we add a polynomial

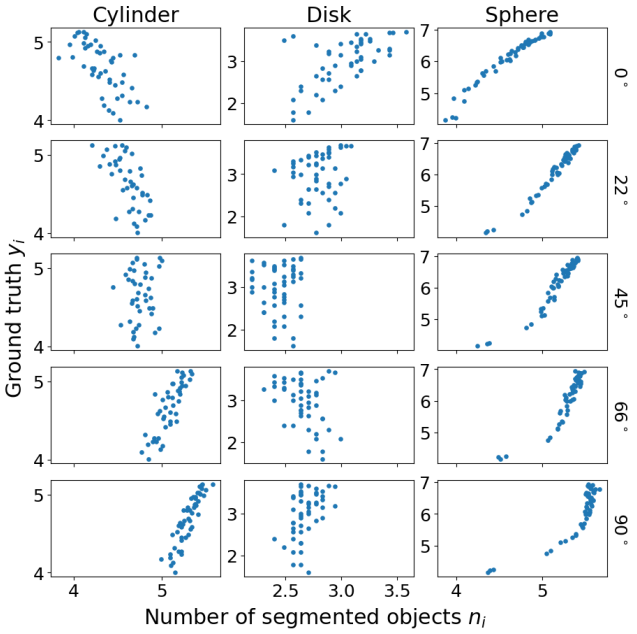

Figure 3: The relationship between the number of segmentation segmented objects n_i and ground truth y_i in a log-log plot. Each row corresponds to a viewing angle shown on the right and each column corresponds to one shape. Only cylinders, disks, and spheres of size S are plotted here. For a fixed shape, y_i can be approximated by polynomials of n_i and the angle a_{mi}. This trend is learned in the transformation stage.

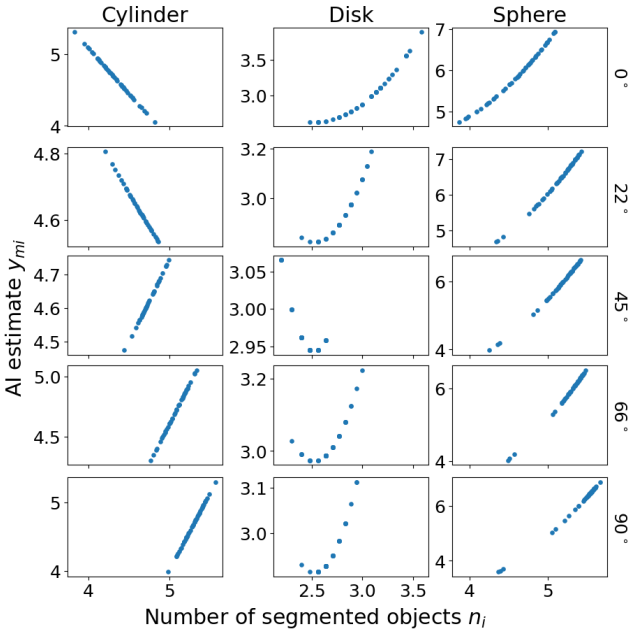

Figure 4: How the number of segmented objects n_i is transformed into the AI estimate y_{mi}. Only cylinders, disks, and spheres of size S are plotted here. For each shape, we learn a family of quadratic functions involving n_i and the angle a_{mi} that aim to capture the relationship between n_i and the ground truth y_i.

regression model to transform the segmentation model's output. The number of objects can range from a few disks to nearly 1000 spheres, which can be problematic for regression and other subsequent steps. Therefore, we take the log of the total number of objects in the jar and the estimates. From now on, these quantities will remain on a log scale.

As was discussed earlier, the angle is a key component in our model, and this regression model is no exception. For an image i with y_i objects in the jar, suppose that the angle is $a_{mi} \in [0, 90]$, and the number of segmented objects given by the segmentation model is n_i. Vector Y_i contains polynomial features constructed using a_{mi} and n_i, further explained below. The regression model is $y_i = Y_i\beta + \epsilon_i$ where β is a parameter and ϵ_i is the error associated with image i. We obtain the ordinary least squares estimator $\hat{\beta}$ from a training data set and then take $Y_i\hat{\beta}$ as the AI estimate y_{mi}.

It remains to be decided how many polynomial features to use. We first need to understand how the segmentation model performs on the images. Figure 3 shows the relationship between the number of segmentation segmented objects and the ground truth broken down by angles. Consider cylinder (S) viewed from different angles in the first column. For every fixed angle, y_i can be captured by a function of n_i, and we can see that the trend is not linear. With degree 2 polynomial features $Y_i = \left[1, \quad a_{mi}, \quad n_i, \quad a_{mi}^2, \quad a_{mi}n_i, \quad n_i^2\right]$, for any fixed angle, we will have $Y_i\beta$ as a quadratic function of n_i, resulting in five quadratic functions for cylinder (S). These should fit the trend we see.

The effect of transformation is in Figure 4. We also experimented with degree 3 polynomial features and noticed signs of overfitting after plotting the polynomials learned. One observation is that for disk (S) viewed from a 90° angle, there is no clear relationship in Figure 3, and the quadratic function cannot approximate a vertical line. But when there is a clear trend, such as sphere (S) viewed from a 0° angle, the model learns to approximate it well. For objects of size L, we also plotted the relationship between n_i and y_i as well as the effect of transformation, and these are in Supplemental materials.

Evaluation

Human Data Set

The human data set is from an unpublished behavioral study, where each participant provided their estimates for the number of objects in some images. There were 205 participants and each was given 30 images, amounting to 6150 trials in total.

In the study, a total of 300 unique images were created by simulating dropping objects into a jar. In each image, the jar can contain varying numbers of cylinders, disks, or spheres. Each comes in two sizes, small (S) and large (L). For cylinders and spheres, the size only affects their radius, and for disks, the size only affects their height. For each jar simulated, there were five images of it viewed from five different angles: 0° (horizontal view), 22°, 45°, 66°, and 90° (vertical view). In summary, for each of the six distinct shapes, 10 separate jars were simulated, and there were five images for each jar, resulting in 300 images. The number of objects in the jar is between 57 and 168 for cylinder (S), 9 and 50 for cylinder (L), 10 and 37 for disk (S), 6 and 27 for disk (L), 314 and 996 for sphere (S), and 249 and 872 for sphere (L).

Due to the noisy nature of human data, we decided to remove outliers. For any image, we take all the estimates given by people who saw it, compute the mean and standard deviation, and keep those within 3 standard deviations of the mean. In this way, extremely small or large estimates for any image were removed. Note that the mean and standard deviation are computed per image, not across all images, since a large value could be appropriate for one image but considered extreme for another image with very few objects. After this preprocessing step, 6001 out of 6150 (97.6%) trials remain.

Additional Images and Overall Setup

For the AI, we have two types of models: one for transformation and one for the data-generating process. We generated additional images to learn the parameters of these models so that those images given to humans would be entirely new to the AI. Details on these additional images can be found in Supplemental materials.

Overall, we performed 5-fold cross-validation. We used 80% of human data to learn the parameters in their data-generating process. For each human trial in the remaining 20%, we paired it with the AI estimates for the same jar across all five possible angles. This would be our test data set. For example, a human trial has a 0° view of a jar. We paired it with the AI estimates for this jar with viewing angles 0°, 22°, 45°, 66°, and 90°. Note that each image was given to multiple participants in the human data set. But in our setup, the AI is paired with one human at a time, and it does not remember that an image had been shown before. Our framework can be adapted if we want the AI to recognize this.

Baselines and Metrics

We compare the performance of our model against the following competitive baselines: average, weighted average, and pick the best. Given two predictions from two agents, the average baseline takes the average of them. This may seem simple but in some circumstances, it outperforms more sophisticated methods (Merkle, Saw, and Davis-Stober 2020).

Next, we introduce the weighted average baseline. For an image i, suppose that $\hat{y}_{hi}$ is the human prediction, $\hat{y}_{mi}$ is the AI prediction, and w is the weight given to the human prediction. The weighted average of these two predictions is $\hat{y}_i = w\hat{y}_{hi} + (1-w)\hat{y}_{mi}$. This baseline finds w that minimizes $\sum_i (y_i - \hat{y}_i)^2$ and uses the optimal weight $\hat{w} = \frac{\sum_i (y_i - \hat{y}_{mi})(\hat{y}_{hi} - \hat{y}_{mi})}{\sum_i (\hat{y}_{hi} - \hat{y}_{mi})^2}$ to compute the combined predictions. The derivation can be found in Supplemental materials. Recall that our model uses a separate set of parameters for each shape. For a fair comparison, this baseline also uses separate weights for each shape. We split the training and test data set by shape. Optimal weight is determined using the training data set for each shape and then applied to the corresponding test data set. This baseline does not take into

account the different angles, as the weight is fixed for each shape.

Furthermore, we have the pick the best baseline. It utilizes angle information and has access to the ground truth in the test data set. The two agents can each view the jar from five angles, so we have 25 possible angle combinations for the human-AI pair. For each of the six shapes and 25 angle combinations, we choose the agent with the smaller MSE to make predictions. For example, suppose the shape is cylinder (S), the human has a 0° view, the AI has a 90° view, and the AI has better performance in that case. Then this baseline will use AI predictions as combined predictions. If we have a model better than this baseline, then the combined predictions of our model will be better than the predictions of both agents, and complementarity will be achieved.

To evaluate the performance, we compute the mean squared error (MSE) of the predictions on the test data set and average over 5-fold cross-validation. Furthermore, we define a metric called relative improvement (RI), which is a percentage that is easier to interpret than numeric values of MSE. We have the following definitions:

- Baseline error: The error metric (MSE in our case) of a baseline model.
- Model error: The same error metric for our model.
- Relative improvement (RI): $\left(\frac{\text{Baseline error} - \text{Model error}}{\text{Baseline error}}\right) \times 100\%$.

For each of the six shapes and 25 angle combinations, we compute the relative improvement (RI). Angle combinations are further divided into two categories: humans and AI have the same viewing angle or different angles. We then take the average of the relative improvement (RI) within each category.

Results

Same Angle Performance

Humans and AI make different types of errors. Figure 5 shows the performance of the human predictions, the AI predictions, as well as the combined predictions produced by the AI. For disks, humans perform better than AI when their viewing angle is 0° since they can manually count the number of disks in the jar. For cylinder (S) and spheres, humans are much worse than the AI, suggesting that humans find it hard to estimate a large number of objects. We can also see that our model combines the two estimates in an effective fashion and outperforms humans and AI in almost all cases. When humans perform much worse than the AI, the combined predictions are dominated by the AI prediction, and the MSE of those two are nearly equal. This can be seen from cylinder (S) and spheres. For the other shapes, the two agents have comparable performance, combining their predictions has a clear advantage over both agents alone.

Overall Performance

Relative improvement (RI) results compared to various baselines are presented in Table 1. First of all, our model

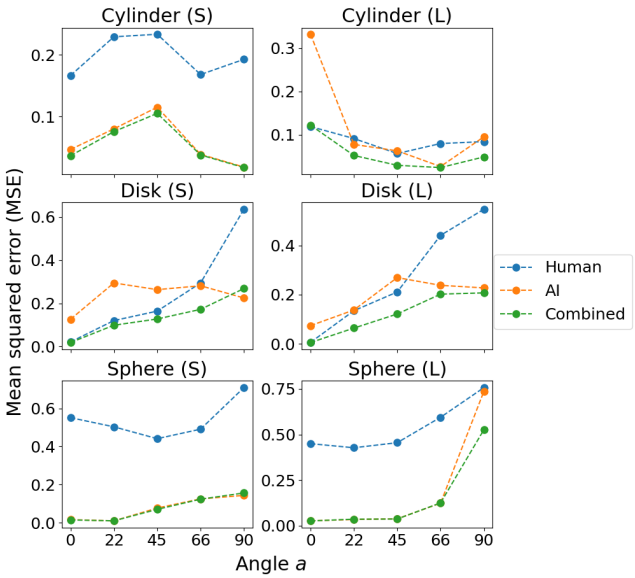

Figure 5: Mean squared error (MSE) of the human predictions, the AI predictions, and the combined predictions when the human and AI have the same viewing angle. The results are divided by the type of object in the jar. When the combined predictions are below the other two markers, complementarity is achieved. Even when this is not the case, the combined predictions are only marginally worse.

is consistently better than the average and the weighted average baselines. These two baselines are unaware of the angle information and cannot adjust the weight when the angle combination changes. A larger RI is usually achieved when the two agents have different views. This suggests that accounting for the information asymmetry in the input sources can lead to more accurate combined predictions. In addition, our model outperforms the pick the best baseline in almost all cases, exhibiting complementarity. This shows that combining across agents is better than delegating tasks to a single agent. Recall that the pick the best baseline has access to the ground truth in the test data set. Without this information, it would not be possible to know which agent to pick in a specific scenario. Even if the agent with better performance is picked, the prediction from the other is completely discarded by this baseline, losing valuable information from this other source. In contrast, our model can incorporate both predictions by leveraging the variance of two agents. One difficult case for our model is sphere (S). From Figure 5 we see that humans are much worse than AI in this case. Our model places little weight on the human predictions, effectively picking the AI, and this explains why it achieves roughly the same performance as the baseline.

Discussion

We introduced a flexible framework that aggregates judgments under information asymmetry, and we evaluated our model using a behavioral data set. The results show that integrating a different source of information can improve the

	Average		Weighted average		Pick the best	
Shape	Same angle	Different angle	Same angle	Different angle	Same angle	Different angle
Cylinder (S)	33.85	34.08	0.97	2.69	7.38	7.98
Cylinder (L)	3.21	10.05	4.21	11.23	21.50	23.99
Disk (S)	17.15	20.22	17.91	20.64	2.12	3.00
Disk (L)	18.48	30.55	19.53	28.16	20.13	16.88
Sphere (S)	60.99	60.50	0.01	2.89	1.36	−1.18
Sphere (L)	52.04	53.80	20.82	27.82	3.73	7.41

Table 1: Relative improvement (RI) in percentage over average, weighted average, and pick the best baselines. S = small; L = large. Each row is the result for a certain shape. "Same angle" columns are averaged across cases where the human and AI have the same viewing angle. "Different angle" columns are averaged across cases where the human and AI have different viewing angles.

overall performance of human-AI teams. Even when the AI can outperform humans on the same task, it can still benefit from a human teammate who approaches the problem from a better vantage point. At the same time, different sources of information are linked together by the common underlying task. When humans and AI have different views, the underlying jar is the same, so they implicitly observe overlapping features. This nuanced setting goes beyond the synthetic data with simple overlapping features (Rastogi et al. 2023).

When humans and AI have the same viewing angle, they incur different errors and the combined predictions can outperform both, suggesting that capability asymmetry can also contribute to complementarity, which was previously explored using behavioral experiments involving image classification (Hemmer et al. 2024). Furthermore, both agents have access to some unique information in our setting, which addresses a limitation in previous work where the AI did not have access to additional information unavailable to humans (Hemmer et al. 2024).

There are several possible directions for future research. For example, although angle is a continuous variable in our model, its effect was not fully explored with only five angles in the data set. Also, modeling characteristics of each individual could be beneficial (Merkle, Saw, and Davis-Stober 2020), but was not utilized here. Most importantly, our framework can be extended to a crowdsourcing setting with multiple humans and multiple AI agents. Previous research investigated a pre-trained classifier and multiple human experts (Showalter et al. 2024). The generalizability of our framework comes from the data-generating process. At its core, one ground truth generates multiple noisy estimates according to some distributions that depend on the agent and the specific scenario. Once we approximate the posterior distribution of the ground truth, we can make predictions. The agents involved can be multiple humans and multiple AI agents, each with their characteristics encoded into the distributions.

In conclusion, our study adds to the existing literature on human-AI collaboration (Rastogi et al. 2023; Hemmer et al. 2024) by providing a framework centered around information asymmetry. We hope that our framework can inspire future work that aims to achieve complementarity when information asymmetry is present and can be leveraged.

References

Anders, R.; Oravecz, Z.; and Batchelder, W. H. 2014. Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, 61: 1–13. Publisher: Academic Press Inc.

Benjamin, D. M.; Morstatter, F.; Abbas, A. E.; Abeliuk, A.; Atanasov, P.; Bennett, S.; Beger, A.; Birari, S.; Budescu, D. V.; Catasta, M.; Ferrara, E.; Haravitch, L.; Himmelstein, M.; Hossain, K. T.; Huang, Y.; Jin, W.; Joseph, R.; Leskovec, J.; Matsui, A.; Mirtaheri, M.; Ren, X.; Satyukov, G.; Sethi, R.; Singh, A.; Sosic, R.; Steyvers, M.; Szekely, P. A.; Ward, M. D.; and Galstyan, A. 2023. Hybrid forecasting of geopolitical events†. *AI Magazine*, 44(1): 112–128. Publisher: John Wiley and Sons Inc.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under Human Assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2611–2620. Number: 03.

Escobedo, A. R.; Moreno-Centeno, E.; and Yasmin, R. 2022. An axiomatic distance methodology for aggregating multimodal evaluations. *Information Sciences*, 590: 322–345.

Hemmer, P.; Schemmer, M.; Kühl, N.; Vössing, M.; and Satzger, G. 2024. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. ArXiv:2404.00029 [cs].

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, 467–474. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9817381-1-6.

Kameda, T.; Toyokawa, W.; and Tindale, R. S. 2022. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, 1(6): 345–357. Number: 6 Publisher: Nature Publishing Group.

Ke, L.; Ye, M.; Danelljan, M.; Liu, Y.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2023. Segment Anything in High Quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934.

Kemmer, R.; Yoo, Y.; Escobedo, A.; and Maciejewski, R. 2020. Enhancing Collective Estimates by Aggregating Cardinal and Ordinal Inputs. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8: 73–82.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. Paris, France: IEEE. ISBN 9798350307184.

Kobayashi, M.; Wakabayashi, K.; and Morishima, A. 2021. Human+AI Crowd Task Assignment Considering Result Quality Requirements. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9: 97–107.

Lahat, D.; Adali, T.; and Jutten, C. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proceedings of the IEEE*, 103(9): 1449–1477. Conference Name: Proceedings of the IEEE.

Liu, C.; Zhong, Y.; Zisserman, A.; and Xie, W. 2022. CounTR: Transformer-based Generalised Visual Counting. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.

Lubars, B.; and Tan, C. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mayer, M.; and Heck, D. W. 2023. Cultural consensus theory for two-dimensional location judgments. *Journal of Mathematical Psychology*, 113. Publisher: Academic Press Inc.

Merkle, E. C.; Saw, G.; and Davis-Stober, C. 2020. Beating the average forecast: Regularization based on forecaster attributes. *Journal of Mathematical Psychology*, 98: 102419.

Merkle, E. C.; and Steyvers, M. 2011. A psychological model for aggregating judgments of magnitude. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6589 LNCS: 236–243. Publisher: Springer, Berlin, Heidelberg ISBN: 9783642196553.

Pinski, M.; Adam, M.; and Benlian, A. 2023. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.

Rastogi, C.; Leqi, L.; Holstein, K.; and Heidari, H. 2023. A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1): 127–139. Number: 1.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 5389–5400. PMLR. ISSN: 2640-3498.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics.

Russakovsky, O.; Li, L.-J.; and Fei-Fei, L. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2121–2131. ISSN: 1063-6919.

Serre, T. 2019. Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, 5: 399–426.

Showalter, S.; Boyd, A. J.; Smyth, P.; and Steyvers, M. 2024. Bayesian Online Learning for Consensus Prediction. In *International Conference on Artificial Intelligence and Statistics*, 2539–2547. PMLR.

Stan Development Team. 2024. Stan Modeling Language Users Guide and Reference Manual.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.

Surowiecki, J. 2004. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. New York, NY, US: Doubleday & Co. ISBN 978-0-385-50386-0. Pages: xxi, 296.

Tan, C. S.; Gupta, A.; and Xu, C. 2022. Are Two Heads Always Better Than One? Human-AI Complementarity in Multi-criteria Order Planning. In *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 0939–0943. Kuala Lumpur, Malaysia: IEEE. ISBN 978-1-66548-687-3.

Taves, E. H. 1941. Two mechanisms for the perception of visual numerousness. *Archives of Psychology (Columbia University)*, 265: 47–47.

Trick, S.; Rothkopf, C. A.; and Jäkel, F. 2023. A normative model for Bayesian combination of subjective probability estimates. *Judgment and Decision Making*, 18: e40. Publisher: Cambridge University Press.

Vaughan, J. W. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research*, 18(193): 1–46.

Yang, Y.; Zhang, L.; Xu, G.; Ren, G.; and Wang, G. 2024. An evidence-based multimodal fusion approach for predicting review helpfulness with human-AI complementarity. *Expert Systems with Applications*, 238: 121878.

Yoo, Y.; Escobedo, A. R.; Kemmer, R.; and Chiou, E. 2024. Elicitation and aggregation of multimodal estimates improve wisdom of crowd effects on ordering tasks. *Scientific Reports*, 14(1): 2640. Number: 1 Publisher: Nature Publishing Group.

Zahedi, Z.; and Kambhampati, S. 2021. Human-AI Symbiosis: A Survey of Current Approaches. ArXiv:2103.09990 [cs].

Zhang, Q.; Lee, M. L.; and Carter, S. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–28. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.

Zhang, Y.; Shang, L.; Zong, R.; Wang, Z.; Kou, Z.; and Wang, D. 2021. StreamCollab: A Streaming Crowd-AI Collaborative System to Smart Urban Infrastructure Monitoring in Social Sensing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9: 179–190.

Zong, R.; Zhang, Y.; Stinar, F.; Shang, L.; Zeng, H.; Bosch, N.; and Wang, D. 2023. A Crowd–AI Collaborative Approach to Address Demographic Bias for Student Performance Prediction in Online Education. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11(1): 198–210. Number: 1.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment Everything Everywhere All at Once. *Advances in Neural Information Processing Systems*, 36.