Wisdom of the Machines: Exploring Collective Intelligence in LLM Crowds

Yashar Talebirad*
University of Alberta
Edmonton, Canada
talebira@ualberta.ca

Ali Parsaee*
University of Alberta
Edmonton, Canada
parsaee@ualberta.ca

Vishwajeet Ohal
University of Alberta
Edmonton, Canada
ohal@ualberta.ca

Amirhossein Nadiri
York University
Toronto, Canada
anadiri@yorku.ca

Csongor Szepesvári
University of Alberta
Edmonton, Canada
csongor@ualberta.ca

Yash Mouje
University of Alberta
Edmonton, Canada
mouje@ualberta.ca

Eden Redman
Network for Applied Technology
Edmonton, Canada
eden@nat.ltd

ABSTRACT

The "wisdom of crowds" phenomenon shows that aggregating independent estimates can yield more accurate predictions than individual guesses. While crowd-sourcing is widely applied, using large language models (LLMs) for collective estimation is largely unexplored. This work investigates how to best form an LLM "crowd" for ambiguous vision-based estimation tasks. We explore two sources of diversity: response diversity, from sampling at various temperatures, and model diversity, from using different LLM architectures. We evaluate these approaches on three vision-based datasets: human height-weight pairs, small objects with known weights, and Amazon products with their prices. Our results show that aggregating deterministic (temperature 0) outputs from a diverse set of models is the most effective strategy, outperforming any single model, as temperature-induced diversity provides no significant benefit while increasing inference costs. The median aggregation of deterministic responses from multiple models outperformed 68% of individual guesses on average, with context providing a significant additional benefit for tasks where it is directly informative (e.g., product titles for price estimation), demonstrating that model diversity is the key to leveraging the wisdom of LLM crowds. By establishing core principles for forming an effective LLM crowd, this work provides a stepping stone for more complex, LLM-driven social simulations.

KEYWORDS

Large Language Models; LLMs; Wisdom of Crowds; Ensemble Methods; Collective Intelligence; Agent-Based Modeling; Vision-Language Models

*Equal contribution

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/MTWR9974

ACM Reference Format:
Yashar Talebirad, Ali Parsaee, Vishwajeet Ohal, Amirhossein Nadiri, Csongor Szepesvári, Yash Mouje, and Eden Redman. 2026. Wisdom of the Machines: Exploring Collective Intelligence in LLM Crowds. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 8 pages. https://doi.org/10.65109/MTWR9974

1 INTRODUCTION

A century after Galton's famous demonstration of a crowd correctly guessing an ox's weight, the "wisdom of crowds" continues to inspire research in social science and machine learning [3, 18]. Large language models (LLMs) offer a modern take on this concept, as instead of being fixed predictors, they generate varied outputs through stochastic processes, each with its own biases, offering different perspectives on a problem [20]. This response diversity allows treating each model call as an independent agent, like an individual in a crowd. Furthermore, our setting falls under what Kameda et al. [8] classify as *combined decisions* within the broader collective-intelligence taxonomy: each model produces an independent estimate with no inter-model communication, and the estimates are aggregated post-hoc. This is distinct from *consensus decisions*, where agents interact before reaching a joint answer.

Modern LLMs can produce multiple outputs in one inference call via sampling methods. These outputs represent diverse predictions, each with a numerical estimate and token probabilities that reflect the model's internal uncertainty [20]. Aggregating outputs from different models, or generating multiple outputs from a single model using various temperatures, lets us explore two distinct sources of diversity. An ensemble of different models provides *model diversity*, where each agent has a unique architecture and training background. Alternatively, sampling a single model repeatedly at varying temperatures generates *response diversity* because of the stochasticity in the sampling process.

The former is analogous to traditional machine learning ensembles, where combining predictions from different models or samples reduces individual errors and improves overall performance [6].

However, unlike classical ensemble methods where base learners are designed to be complementary [6], LLMs produce unstructured, stochastic outputs, making it unclear how best to combine them. Moreover, the relationship between individual sampling variability and collective accuracy in generative language models remains underexplored [9].

In this work, we ask: *What is the most effective way to generate and aggregate diverse outputs from LLMs to achieve higher predictive accuracy?* We address this by formalizing LLM aggregation as a computational extension of crowd wisdom and comparing model diversity versus response diversity. In this work, we find that aggregating deterministic (temperature 0) outputs from a diverse model ensemble generally yields comparable to superior estimates compared to aggregating repeated samples at varied temperatures from a single model. Both methods generally outperform standard sampling, including a single model deterministic sample. Both model-diverse and response-diverse estimates sit along an error-cost Pareto frontier, though response-diverse estimates rely on specific conditions to be efficient.

As LLMs are increasingly used for tasks that require numerical judgment (e.g., cost estimation, forecasting, and assessment), understanding how to best aggregate their output is of practical importance. Our work suggests that combining outputs from different models can yield superior numerical estimates. This ensemble approach may offer two potential advantages: it could reduce reliance on computationally expensive large models by leveraging multiple smaller models, and it may mitigate the systematic biases inherent to any single LLM by combining complementary predictions. We evaluated our approach on three vision-based estimation tasks using datasets for human height and weight, object mass from images, and product prices from Amazon listings. These datasets allow us to focus on predicting numerical values from images, sometimes with additional context. Section 3.1 provides more details on our datasets and methodology.

Surowiecki [18] defines several criteria for crowd wisdom. In Section 3.1, we discuss these criteria and how our experiments are designed to meet them. We used several vision-enabled LLMs, including Qwen2 Vision Language 72B Instruct and Llama 3.2 Vision 11B (from the Together API[1]), and GPT-4o-mini[2]. To investigate response diversity, we varied the temperature across five settings (0.2, 0.4, 0.6, 0.8, and 1.0) in our initial experiments, treating each API call as an independent agent. This mirrors the conditions of Galton's experiment. By combining these independent guesses, we aimed to find the best strategy for a collective prediction that is more accurate than any single response.

On average, the aggregate outperforms 68% of individual responses, with context providing additional benefit for tasks where it is directly informative (e.g., product titles for price estimation). We also explore different aggregation methods and the impact of additional context, discussing their relative performance and implications.

[1]https://docs.together.ai/docs/vision-overview
[2]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

2 RELATED WORK

The combination of independent numerical estimates improving prediction accuracy is a widely recognized principle in social science and machine learning. In the early 20th century, Galton's ox-weight guessing experiment [3] showed that the median of diverse estimates approximates the true value. This "wisdom of crowds" effect [18] depends on diverse and independent estimates, which help cancel out errors. Simoiu et al. [16] found that the median outperforms 65% of individual guesses, confirming the effect in humans.

This concept has influenced ensemble methods in machine learning. Techniques like bagging [1] and boosting [2] combine multiple models to reduce variance and prevent overfitting, often improving accuracy. In computer vision, deep ensembles (e.g., averaging predictions from several deep residual networks) have achieved state-of-the-art performance [7].

Similarly, in NLP, aggregating outputs from various language models improves results. Recent research on large language models (LLMs) shows that some reasoning abilities emerge only in sufficiently large models [20]. Self-consistency decoding [19] combines multiple outputs to improve reliability and accuracy, suggesting that ensemble methods can reveal latent capabilities. Lau et al. [9] and Guo et al. [5] explore how varying prompts elicits diverse reasoning outputs. Lau et al. [9] vary prompt wording to examine problems from different angles, while Guo et al. [5] use multiple prompts to reduce issues like reasoning hallucinations. Both studies show that careful prompt design enhances the ensemble effect, leading to more reliable predictions. In a setting without ground truth, Mousavi Davoudi et al. [11] show that agreement among diverse LLMs can signal the reliability of outputs and the quality of generated questions.

Pratt et al. [13] ask whether forecasting strategies can improve LLM decision-making. This research aligns with agent-based models by Gao et al. [4], where LLMs act as autonomous agents in simulations. Models simulating human behavior, such as in Park et al. [12], show that artificial agents can mimic social dynamics. Li et al. [10] examine if LLMs can understand others' beliefs to encourage collaboration, while Shi et al. [15] propose that agent interactions can reduce reasoning errors.

Schoenegger et al. [14] compare LLM ensemble forecasts with aggregated human predictions in a forecasting tournament. They used an ensemble of twelve LLMs for binary predictions on 31 questions, comparing the result to 925 human forecasters over three months. Their analysis shows aggregated LLM predictions outperform a no-information benchmark and are statistically indistinguishable from human forecasts (within medium-effect-size equivalence bounds). They also show forecasting accuracy improves when models see the median human prediction, but simply averaging human and machine outputs is best. While their work focuses on binary forecasts, which are a useful benchmark for the wisdom-of-the-crowd effect, our study extends this paradigm to continuous estimation tasks using vision-enabled LLMs. Numerical guesses, like weight or cost, have more direct real-world applicability than binary predictions. By aggregating continuous outputs from LLMs at different temperatures, our approach attempts to utilize prediction variability for more accurate and robust estimates.

This work aligns with the findings of Schoenegger et al. [14] by demonstrating that ensemble methods are also effective for complex tasks with visual input and continuous-valued outputs. While ensemble learning, prompt diversity, and agent-based modeling are often studied independently, our approach combines all three, treating LLM configurations as heterogeneous agents whose collective output can be systematically combined.

The next section describes our experimental methodology for simulating this ensemble behavior and evaluating its predictive accuracy across different datasets.

3 METHODOLOGY

3.1 Experimental Setup

3.1.1 Datasets and Data Selection. We used three datasets, randomly sampling 100 items from each with a fixed seed (42) for reproducibility. First, from Kaggle's "Height-Weight Images" dataset [3], we used photos of people with their known weight (lbs) and height (feet, inches). Second, from the Image2Mass dataset [17], we used photos of small objects with their weight (lbs, converted to grams) and dimensions (inches). Third, from an Amazon Canada listings dataset by Asaniczka[4], we used product images and their prices (CAD). These domains were chosen to span a range of perceptual difficulty, output distribution, and semantic context. Estimating human body weight from appearance is a socially grounded, continuous task with high visual ambiguity as inferring object mass from images requires physical intuition about material and geometry. Furthermore, predicting product prices from Amazon listings introduces semantic and market-driven variability less tied to direct visual cues. While this selection is not exhaustive, it allows us to investigate whether crowd-aggregation benefits generalize across meaningfully distinct vision-based estimation tasks.

3.1.2 Models and Configuration. We used three vision-language models (using the available model versions in June 2025):

- via OpenAI API: **GPT-4o-mini**
- via Together API: **Qwen2-VL-72B-Instruct** and **Llama-3.2-11B-Vision-Instruct-Turbo**

3.1.3 Prompts Used. Table 1 provides the exact prompts used in our experiments.

3.1.4 Experimental Parameters. For the initial two datasets, we tested five temperature settings for each of the three models: 0.2, 0.4, 0.6, 0.8, and 1.0. Each configuration was repeated 15 times per image, resulting in 225 total API calls per image (3 models × 5 temperatures × 15 repetitions). Based on our finding that temperature adds more noise than signal (see Section 4), our Amazon Price experiments used only temperature 0. For this dataset, each model was queried once per image. All other parameters were held constant: max_tokens=10 and top_p=1.0.

3.1.5 Task Definition and Wisdom of Crowds Criteria. The task was to estimate a numerical value (weight or price) from an image.

[3]https://www.kaggle.com/datasets/virenbr11/height-weight-images
[4]https://www.kaggle.com/datasets/asaniczka/amazon-canada-products-2023-2-1m-products

Table 1: Prompt templates used for each dataset and context condition. Variables in brackets were replaced with actual values during experiments.

Dataset	Without Context	With Context
Height-Weight	"Based solely on the image, give your best numeric estimate of the weight (in lbs) of the person. Output only the number and nothing else."	"Based on the image and the additional information that this person is [HEIGHT] tall, give your best numeric estimate of their weight (in lbs). Output only the number and nothing else."
Image2Mass	"Based solely on the image, give your best numeric estimate of the weight (in grams) of the object. Output only the number and nothing else."	"Based on the image and knowing that the object has dimensions [DIMENSIONS] inches, give your best numeric estimate of the weight (in grams) of the object. Output only the number and nothing else."
Amazon Price	"Based solely on the image, give your best numeric estimate of the price (in CAD) of the product. Output only the number and nothing else."	"Based on the image and the product title '[TITLE]', give your best numeric estimate of its price (in CAD). Output only the number and nothing else."

The "wisdom of crowds" relies on several criteria for a group to produce accurate collective judgments. As identified by Surowiecki [18], these are:

(1) **Diversity:** Each individual contributes unique insights. This variance in opinion helps to counterbalance errors and biases, improving collective accuracy.
(2) **Independence:** Judgments must be independent. Uncorrelated errors tend to cancel out when aggregated, making the collective estimate more accurate.
(3) **Decentralization:** Decision-making should be decentralized, allowing individuals to use their own knowledge.
(4) **Aggregation Mechanism:** A mechanism is needed to aggregate individual judgments into a collective decision, from simple averaging to more complex weighted combinations.

These conditions allow for a robust and accurate collective decision that can exceed individual capabilities [18], mirroring principles of decentralized social systems.

3.1.6 Experimental Design. To meet these criteria, we experimented with different LLMs in various settings (Figure 1). This allowed us to

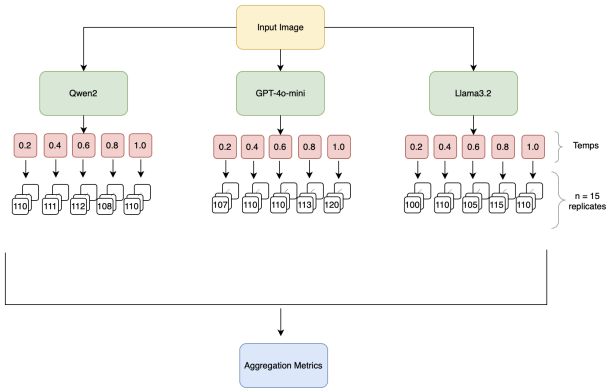

Figure 1: Architecture of the initial LLM ensemble used for the Height-Weight and Image2Mass datasets. The Amazon Price experiments followed a simpler approach.

compare two primary sources of diversity: **model diversity**, from using an ensemble of different models, and **response diversity**, from generating multiple outputs from a single model at various temperatures. Temperature controls output randomness (lower is more deterministic). This will allow us to compare whether a crowd of diverse models is wiser than a crowd of diverse responses from a single model. We treated each API call as an independent agent with no shared context or communication between calls, in an attempt to mimic collecting diverse, independent, and decentralized guesses from a crowd. In Section 3.2, we consider different aggregation methods to satisfy the fourth criterion.

We hypothesize that different LLM architectures may provide meaningful diversity due to their distinct training procedures, architectures, and data compositions. The three models we selected (GPT-4o-mini from OpenAI, Qwen2-VL-72B from Alibaba, and Llama-3.2-11B from Meta) were developed by different organizations with varying model sizes (11B to 72B parameters), training objectives, and likely different training data distributions. These architectural and training differences could lead to systematically different biases in estimation tasks, as suggested by the empirical patterns in Figure 2. While these models may share some common training data from public sources, their distinct processing architectures and training methodologies likely result in different internal representations and systematic biases. Our empirical results suggest that these models exhibit complementary error patterns that tend to cancel when aggregated, supporting the diversity assumption for our specific task domain.

Finally, we aggregated the outputs to see if the collective estimate could outperform individual predictions. Comparing these aggregates to the ground truth allowed us to assess the collective and predictive potential of LLM ensembles. Human crowds often use context to improve their predictions. We simulated this by giving LLMs extra context, hypothesizing it would improve accuracy. For the Height-Weight and Image2Mass datasets, we provided context by including the person's height or the object's dimensions in the prompts. We then compared predictions with and without context to assess their impact. For the Amazon Prices dataset, we gave the product title in the prompt.

3.2 Aggregation and Weighting Methods

We combined independent outputs from multiple LLM calls for a robust aggregate estimate. We considered two main unweighted aggregation methods:

(1) **Mean:** The arithmetic mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sensitive to outliers).
(2) **Median:** The median, the middle value of sorted predictions (robust to outliers).

We also explored weighting predictions by token-level confidence (log-probabilities), but our initial analysis showed no significant benefit over unweighted methods. We therefore focus on the mean and median in our main analysis. All aggregation methods are blind to the ground truth, combining only the LLM outputs.

3.3 Ranking

We assess performance using a ranking mechanism similar to Simoiu et al. [16]. For each image, we rank the aggregate's prediction error among the individual prediction errors from the same experimental condition (i.e., same context setting). This rank percentile shows the fraction of individual predictions the aggregate outperforms (a lower percentile means fewer individual predictions were better). For statistical significance, we use a one-sided paired t-test on the rank percentiles for each image.

3.4 Mathematical Model

Consider an entity E with a d-dimensional attribute vector $\boldsymbol{\theta} \in \mathbb{R}^d$. A digital image of it, $P(E)$, is shown independently to m LLM agents at temperature $T = 0$.

Each agent A_i produces an estimate

$$\hat{\boldsymbol{\theta}}_i = f_i\big(P(E)\big) = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ denotes the idiosyncratic error of agent i.

We then form a coordinate-wise median aggregation:

$$\tilde{\theta}_j = \text{median}\big\{\hat{\theta}_{1j}, \hat{\theta}_{2j}, \ldots, \hat{\theta}_{mj}\big\}, \quad j = 1, \ldots, d. \tag{2}$$

This median aggregation can be formally characterized as a Fréchet mean[5], which generalizes the concept of centrality in metric spaces by minimizing the sum of distances to observed points. Specifically, for each coordinate j using the Euclidean distance metric, the median is the solution to:

$$\tilde{\theta}_j = \arg\min_{x \in \mathbb{R}} \sum_{i=1}^{m} |x - \hat{\theta}_{ij}|. \tag{3}$$

The choice of distance metric and aggregation function is critical in determining the aggregation behavior. For continuous real-valued outputs using the Euclidean distance metric $d(x, y) = |x - y|$, different aggregation methods correspond to different powers α in the objective function $\sum_{i=1}^{m} d(x, \hat{\theta}_{ij})^{\alpha}$: the median minimizes the sum of distances ($\alpha = 1$), while the arithmetic mean minimizes the sum of squared distances ($\alpha = 2$). For discrete categorical outputs, majority voting can be viewed as a Fréchet mean using the discrete metric (0-1 loss), where $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ otherwise.

[5] https://en.wikipedia.org/wiki/Frechet_mean

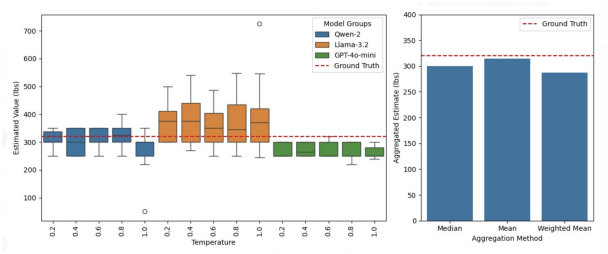

Figure 2: An example from the Height-Weight dataset showing different model biases. Qwen-2 and GPT-4o-mini tend to underestimate weight, while Llama-3.2 overestimates. The median of the estimates is closer to the true value than any single model's.

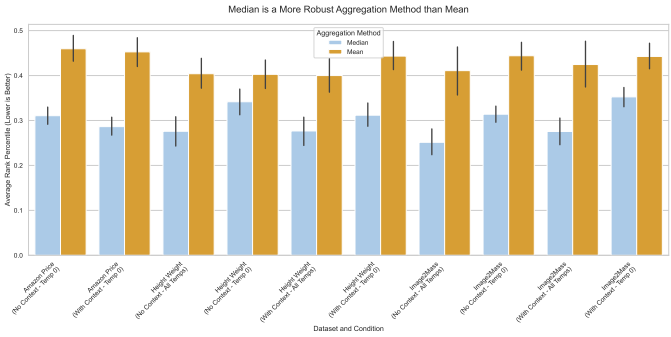

Figure 3: Median vs. Mean aggregation performance. The median consistently achieves a lower (better) average rank percentile, confirming its robustness.

The robustness of the median ($\alpha = 1$) to outliers makes it particularly suitable for aggregating LLM predictions, where individual estimates may occasionally be far from the true value.

We assume each f_i is a distinct world model (due to different architectures, data, and biases), so their error vectors $\boldsymbol{\varepsilon}_i$ have zero median and are weakly correlated ($\mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) \approx 0$ for $i \neq j$). This assumption best applies when aggregating across diverse model architectures, which our results show is most effective.

4 RESULTS AND DISCUSSION

Our findings show that: 1) median aggregation is more effective than the mean; 2) aggregates are better than individuals; 3) model diversity enables robustness; 4) model diversity is efficient; and 5) context benefit is task-dependent. We will now go over each of these statements.

Median aggregation is more effective than the mean. Our first analysis confirmed that the median is superior to the mean for aggregation. Estimation tasks are prone to outliers, causing the mean to perform poorly as an aggregation method. Our experiments (3) show that the median's rank percentile was significantly lower (better) than the mean's across all datasets and conditions ($p < 0.001$ for all testable comparisons). We therefore focus on median-based approaches.

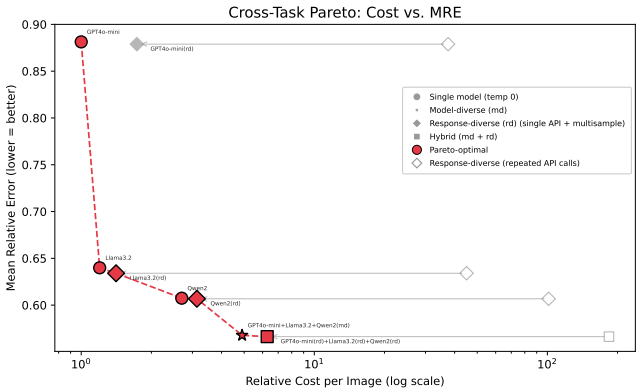

Figure 4: Cost-MRE Pareto frontier plot. Relative cost vs. mean relative error (MRE) for different estimates. The red dashed line represents the Pareto frontier. Model diversity drives the main MRE improvement; single-model temp-0 and response-diversity points are close.

Aggregates are better than individuals. To test for a "wise" crowd, we checked if our best method's rank percentile ('Median - All Models - Temp 0') was significantly below 0.5 (the expected rank of an average individual). A one-sample t-test confirms this with high confidence ($p < 0.0001$) across all conditions. Table 2 and Figure 6 summarize this result. This confirms that the aggregate is not just better than a few noisy individuals but is statistically superior to the typical individual guess, which shows a wisdom of the crowd effect. On average, our method outperforms 68% of individual responses.

Model diversity enables robustness. Figure 2 qualitatively reinforces the hypothesis that aggregation can cancel biases found in individual models to achieve a superior estimate. As shown in Figure 5, the "All Models" median aggregate provides a robust estimate across all datasets and conditions. While individual models occasionally achieve lower error, no single model is consistently best across all tasks. The ensemble provides a reliable strategy that does not require identifying the best model in advance.

Model diversity is efficient. Figure 4 shows the resulting Pareto frontier of relative cost vs. mean relative error (MRE). Single-model temp-0 and response-diverse points are close together: adding temperature-induced diversity to one model yields only marginal MRE gains at a higher cost. The dominant gain comes from *model* diversity. The full 3-model median (temp-0 or hybrid) reaches the best MRE on the frontier. Hollow points represent a naive approach to response-diverse sampling where the image is passed in 75 separate API calls, whereas the filled points represent loading the image a single time, then repeatedly sampling to generate different estimates. Taking the cost of GPT-4o-mini as a baseline of 1 unit cost, we determined the relative costs through API pricing at the time of implementation. It is notable that without the more efficient implementation of response-diversity, these samples do *not* lie along the Pareto frontier at all.

Context benefit is task-dependent. To test whether context improves aggregate accuracy, we compared the mean relative error

($|\hat{y}-y|/|y|$) of the aggregate (Median of all models, Temp 0) with and without context using a paired t-test. Context significantly reduced the aggregate's mean relative error for the Amazon Price dataset (from 88% to 45%, $p = 0.025$), where the product title is highly informative. However, context did not yield a statistically significant improvement for height-weight ($p = 0.13$) or image2mass ($p = 0.15$). We hypothesize that this difference arises because the product title resolves the primary source of uncertainty for price estimation, namely the identity of the product, which the image alone cannot reliably convey. In contrast, for height-weight and image2mass, the provided context (height or dimensions) addresses secondary factors such as scale, while the dominant sources of uncertainty (body composition or material density) remain unresolved. This suggests that supplementary context could be most valuable when it covers the dominant source of estimation uncertainty.

Dataset	Without Context	With Context
Height-Weight	0.342	0.312
Image2Mass	0.314	0.353
Amazon Price	0.311	0.286
Average	**0.322**	**0.317**

Table 2: Average rank percentile of the best aggregate method (Median of All Models, Temp 0), ranked against individual predictions from the same experimental condition. In all conditions, the aggregate significantly outperforms the expected rank of an average individual (0.5), with p < 0.0001 for a one-sample t-test.

5 LIMITATIONS AND FUTURE WORK

Our study has several limitations that suggest directions for future research. First, our analysis uses three datasets with small sample sizes (100 instances each). Larger, more diverse datasets are needed to generalize our conclusions. Second, our experiments only cover estimating weight and price. As Simoiu et al. [16] noted, crowd performance varies by task. More broadly, all three of our datasets involve vision-based numeric estimation from images. While we deliberately selected domains that differ in perceptual difficulty and data distribution, any generalization beyond computer-vision scenarios remains speculative. However, our mathematical aggregation framework (Section 3.4) is domain-agnostic in principle, which motivates future work on non-visual and mixed-modality estimation tasks, as well as forecasting or subjective judgments. Third, we lack a direct comparison to a human baseline for these visual tasks, which would provide valuable context. Fourth, our study is limited to three LLM architectures. While these models demonstrate complementary biases that improve aggregate performance, including additional diverse architectures could further enhance accuracy and provide more robust estimates. The optimal ensemble size and the marginal benefit of adding models remain open questions for future investigation. Fifth, while our approach suggests potential computational advantages by using ensembles of smaller models rather than single large models, we have not conducted a comprehensive analysis of the computational trade-offs in terms of FLOPs, latency, memory requirements, and overall resource consumption. A more thorough study is needed to quantify whether model ensembles are indeed more efficient than larger individual models for specific deployment scenarios. Finally, while our goal is to improve estimation reliability, any method that enhances automated numeric judgment at scale can be misused. Our work does not involve training new models and operates solely on inference-time aggregation of existing API outputs, but responsible application, nonetheless, requires domain-specific oversight.

In future work, we will investigate how sampling parameters (e.g., top-p, top-k) affect output diversity and aggregate accuracy. We also aim to use token-level distributions (entropy, perplexity) for more refined aggregation weighting. Additionally, we plan to explore adaptive weighting schemes that account for each model's systematic biases and task-specific performance patterns, potentially improving aggregate accuracy by assigning higher weights to models that demonstrate lower bias for particular estimation domains. We will also explore dynamic multi-agent interactions where agents adjust predictions based on cues and peer outputs, simulating social learning. This would allow for modeling of influence and belief propagation as in human social networks, enabling large-scale experiments on collective intelligence. Such simulations could systematically test variables like communication topology and information cascades at a scale and with a level of control that is infeasible in human studies. We also plan to use a Fermi-inspired estimation[6] strategy, using chain-of-thought prompting to make LLMs decompose complex tasks into components (e.g., material, dimensions). Estimating and combining these components may enhance final prediction accuracy.

Finally, a promising direction is extending our mathematical framework beyond numerical outputs to a general-purpose aggregation system for diverse output types. Our current work focuses on aggregating scalar predictions using median aggregation. However, many practical applications require aggregating structured outputs such as probability distributions and rankings, which necessitate mapping these outputs to appropriate mathematical spaces with suitable distance metrics. Aggregating *probability distributions* is motivated by the observation that simply selecting the most probable outcome from each model discards valuable information about what each model thinks regarding alternative choices. Aggregating full distributions could improve decision-making in classification tasks and scenarios where models express varying degrees of confidence across multiple options. For *rankings*, we aim to develop methods for aggregating ordered preferences when each model produces a ranked list of items, which has applications in tasks where models propose multiple solutions with implicit or explicit priority orderings. The overarching goal is to establish a unified Fréchet mean-based aggregation framework applicable across multiple domains, enabling wisdom-of-crowds benefits for any task where diverse model outputs can be meaningfully combined.

6 CONCLUSION

Our work shows that "wisdom of the crowds" principles apply to LLM ensembles for vision-based estimation tasks. We find that the

[6] https://en.wikipedia.org/wiki/Fermi_problem

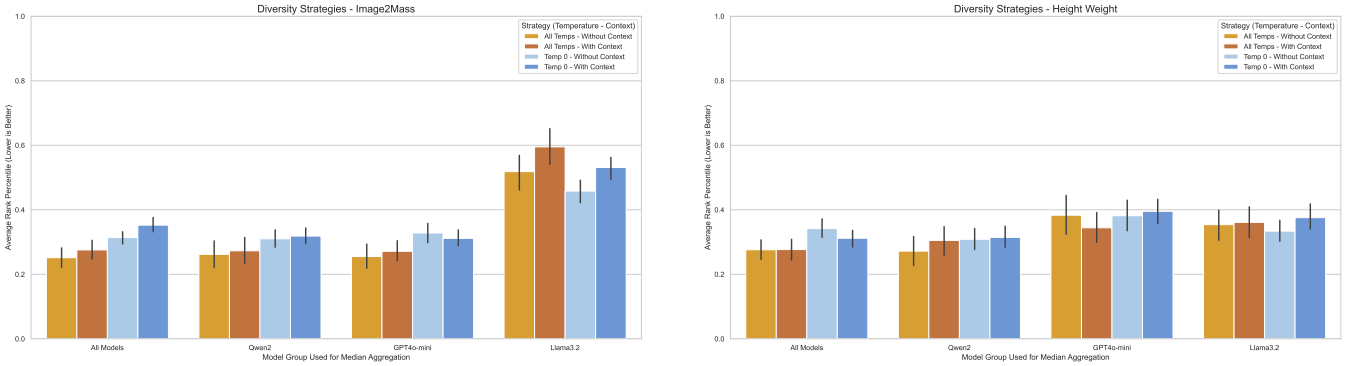

Figure 5: Diversity strategies. Median aggregation performance across model groups and strategies. Aggregating "All Models" (leftmost) performs well, and Temperature 0 (blue) is comparable to All Temperatures (orange), with no statistically significant difference between them. This plot shows results for the two datasets where temperature was varied; the Amazon Price dataset, which only used Temperature 0, is omitted as it does not have an "All Temps" condition to compare against.

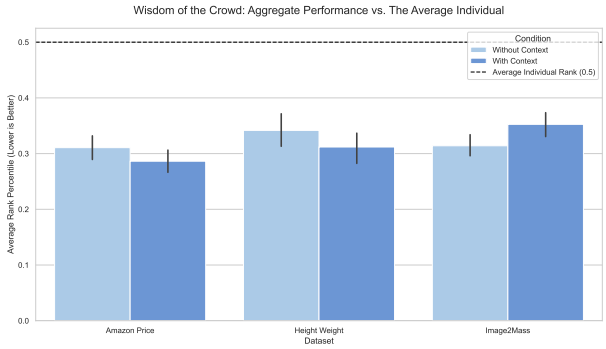

Figure 6: The aggregate is wiser than the average individual. Performance of our best method (Median of All Models, Temp 0). Bars show the average rank percentile (95% CI). The dashed line at 0.5 is the expected rank of an average individual. The aggregate is significantly better in all cases.

source of diversity is critical, as our results demonstrate that aggregating deterministic (temperature 0) outputs from diverse models (model diversity) is a more robust and cost-effective strategy than generating multiple outputs from a single model using temperature sampling (response diversity). This is supported by two of our main findings: 1) the multi-model aggregate provides robust performance across all tasks, without requiring identification of the best individual model in advance, and 2) temperature-induced diversity provided no significant improvement while increasing inference costs. Therefore, the most practical strategy is a committee of diverse "expert" models, each giving its most confident estimate. This approach may also reduce dependence on computationally expensive large models and mitigates the systematic biases inherent to any single LLM by leveraging complementary error patterns across models. By grounding our approach in the Fréchet mean framework, we lay the groundwork for developing general-purpose aggregation methods that could extend beyond numerical predictions to diverse applications including classification, ranking, and decision support across multiple domains.

ACKNOWLEDGMENTS

We thank Network for Applied Technology (NAT)[7] for their funding and resources in support of this work.

REFERENCES

[1] Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140.
[2] Yoav Freund and Robert E. Schapire. 1996. Experiments with a New Boosting Algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*. 148–156.
[3] Francis Galton. 1907. Vox populi. *Nature* 75 (1907), 450–451.
[4] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.
[5] Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Shimin Tao, Hengchao Shang, Zongyao Li, Shaojun Li, Jinlong Yang, Zhanglin Wu, Zhiqiang Rao, et al. 2024. M-ped: Multi-prompt ensemble decoding for large language models. *arXiv preprint arXiv:2412.18299* (2024).
[6] Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 12, 10 (1990), 993–1001.
[7] 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*. 770–778.
[8] Tatsuya Kameda, Wataru Toyokawa, and R. Scott Tindale. 2022. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology* 1 (2022), 345–357.
[9] Gregory Kang Ruey Lau, Wenyang Hu, Diwen Liu, Jizhuo Chen, See-Kiong Ng, and Bryan Kian Hsiang Low. 2024. Dipper: Diversity in prompts for producing large language model ensembles in reasoning tasks. *arXiv preprint arXiv:2412.15238* (2024).
[10] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701* (2023).
[11] Seyed Pouyan Mousavi Davoudi, Amin Gholami Davodi, Alireza Amiri-Margavi, Alireza Shafiee Fard, and Mahdi Jafari. 2025. Collective reasoning among LLMs: A framework for answer validation without ground truth. *arXiv preprint arXiv:2502.20758* (2025).
[12] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.
[13] Sarah Pratt, Seth Blumberg, Pietro Kreitlon Carolino, and Meredith Ringel Morris. 2024. Can Language Models Use Forecasting Strategies? *arXiv preprint arXiv:2406.04446* (2024).

[7] https://www.nat.ltd

[14] Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances* 10, 45 (2024), eadp1528.

[15] Jinxin Shi, Jiabao Zhao, Xingjiao Wu, Ruyi Xu, Yuan-Hao Jiang, and Liang He. 2025. Mitigating reasoning hallucination through Multi-agent Collaborative Filtering. *Expert Systems with Applications* 263 (2025), 125723.

[16] Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. Studying the "wisdom of crowds" at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 171–179.

[17] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. 2017. image2mass: Estimating the mass of an object from its image. In *Conference on Robot Learning*. PMLR, 324–333.

[18] James Surowiecki. 2004. *The Wisdom of Crowds.* Anchor Books.

[19] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[20] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).