Probing LLM World Models: Enhancing Guesstimation with Wisdom of Crowds Decoding

Anonymous ACL submission

Abstract

Guesstimation, the task of making approximate quantity estimates of a physical object or an event, is a common real-world challenge. However, it has largely been overlooked in large language model (LLM) research. We introduce three guesstimation datasets, *MARBLES*, *FUTURE*, and *ELECPRED*. These datasets include guesstimation ranging from concrete object estimation (e.g., how many marbles can fit in a one-cup measuring cup) to abstract scenario predictions such as predicting the 2024 U.S. presidential election result. Inspired by the social science concept of the "*Wisdom of Crowds*" (WOC)—taking the median from estimates from a crowd, which has proven effective in guesstimation, we propose the "WOC decoding" strategy for LLM guesstimation. We replicate prior findings that WOC improves human guesstimation accuracy and show that LLMs exhibit a similar WOC effect. The success of LLMs in guesstimation suggests they possess some level of a "world model" necessary for guesstimation. Moreover, the WOC decoding method improves LLM guesstimation accuracy more efficiently than other decoding methods, such as self-consistency. These results highlight the value of the WOC decoding strategy for LLMs and position guesstimation as a probe for evaluating LLMs' world model. As LLMs' world model is a fundamental prerequisite for many real-world tasks (e.g., forecasting and human-AI teaming), our findings have broad implications for the AI community.

1 Introduction

Daily life often requires us to estimate uncertain quantities, from the crowd size at a political event to the weight of a turkey needed for a Thanksgiving dinner. In human populations, such "guesstimation" scenarios often exhibit *wisdom of crowds* (WOC) effects: in a random sample of estimates, the median lies closer to the ground truth than most individual guesses (Galton, 1907; Yu et al., 2018). WOC phenomena are thought to rely on a *world model*—a conceptual understanding of the world that supports estimation and decision-making. For instance, when estimating the number of jelly beans in a jar (Surowiecki, 2005), people may rely on an implicit understanding of the typical size, shape, and firmness of jelly beans, and the shape, volume, and rigidity of the jar. Even for more abstract scenarios, people may also rely on general world-knowledge; for instance, when estimating the number of people requiring food stamps in Chicago, their guesses may reflect general knowledge/beliefs about poverty rates, accessibility of government programs, characteristics of large midwestern cities, etc.

Here we assess whether contemporary large language models (LLMs) exhibit WOC phenomena similar to those observed in human populations. LLMs are crowds unto themselves: they are trained on vast amounts of linguistic and other data generated and tuned from crowds of individual human users. Thus, multiple samples of responses from a single model may be akin to asking multiple users from a human population the same question, in which case the median of model responses might closely approximate the ground truth.

To systematically study guesstimation and WOC effects in LLMs, we created three guesstimation datasets: *MARBLES*, *FUTURE*, and *ELECPRED*. The MARBLES dataset involves estimating the number of physical objects (e.g., marbles, coins) that can fit into different containers (e.g., one-cup dry-ingredients measuring cup), requiring reasoning based on real-world physical properties. On the other hand, FUTURE and ELECPRED datasets involve guesstimation in more abstract scenarios - predicting future real-world events like population growth, economic trends, or 2024 U.S. presidential election results, all of which require reasoning based on real-world knowledge such as demographics, economic conditions, and political trends.

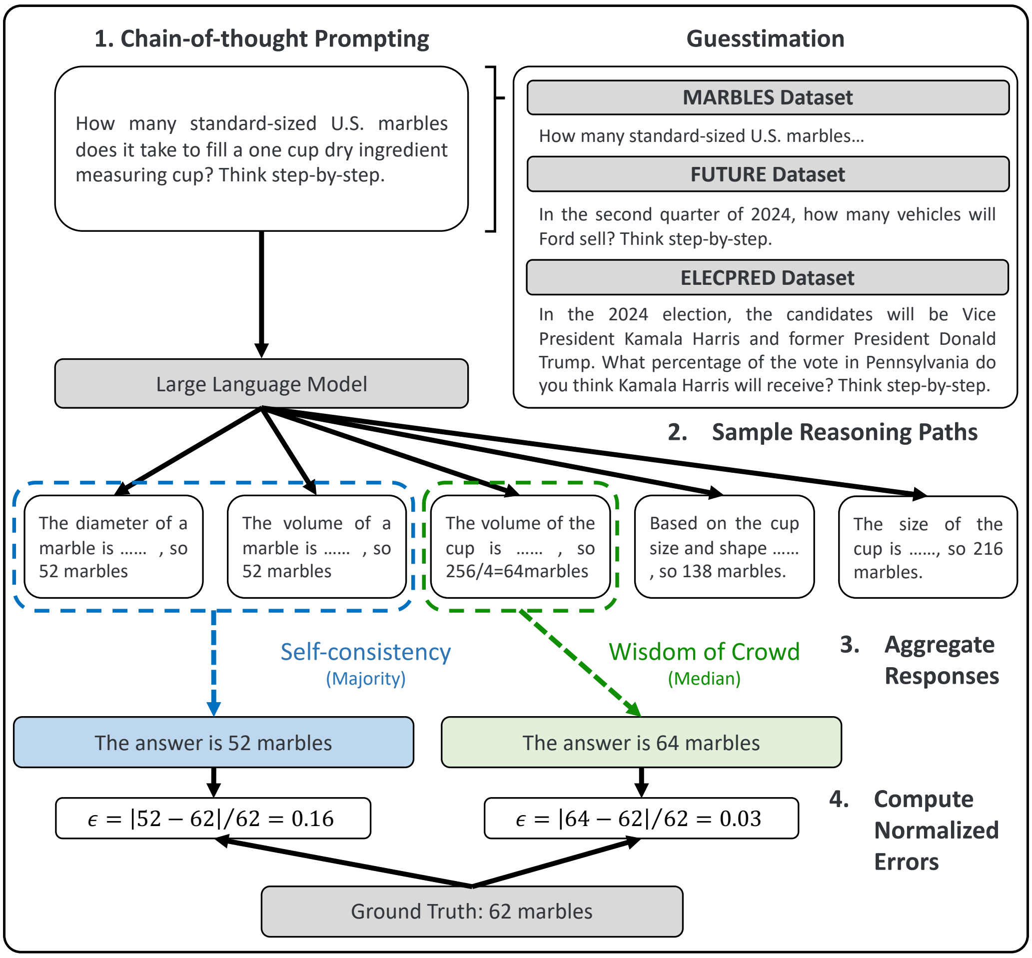

Figure 1: The steps of LLM guesstimation through self-consistency decoding method and wisdom of crowd (WOC) decoding method.

In all experiments, the guesstimation questions were provided in natural language to the LLMs. To quantify the WOC effect in each case, we took the normalized error: the absolute difference between the median guess and the ground truth divided by the ground truth. The more these error terms are reduced with increasing size of the crowds, the greater the WOC advantage relative to an individual guesser. We further compared the LLM WOC behavior with the *self-consistency* decoding strategy, which samples model behavior many times and returns the majority vote among the samples, rather than the median as WOC. Prior work has suggested that self-consistency can improve model reasoning behavior (Wang et al., 2023). In addition, we also conducted a human experiment and replicated previous findings about WOC in human crowds.

Our results demonstrate the effectiveness of WOC decoding in guesstimation tasks in both humans and LLMs. We showed that WOC decoding outperformed self-consistency and greedy decoding across both concrete and abstract guesstimation datasets (including 2024 U.S. presidential election prediction) and achieved greater accuracy with fewer samples. In sum, we proposed guesstimation as a method to probe LLMs' world models, and showcased that we can apply WOC, a social science-inspired decoding strategy, to reach the best guesstimation performance. Our findings have broader implications for real-world applications such as forecasting and human-AI teaming, which rely on an accurate world model. In sum, we introduce guesstimation as a new task that is very common in real world but has been over-looked by the AI community.

2 Methods and Experimental Setup

2.1 Guesstimation Datasets

MARBLES Dataset The *MARBLES* dataset consists of 15 guesstimation questions, involving five different containers (a one-cup dry ingredient mea-

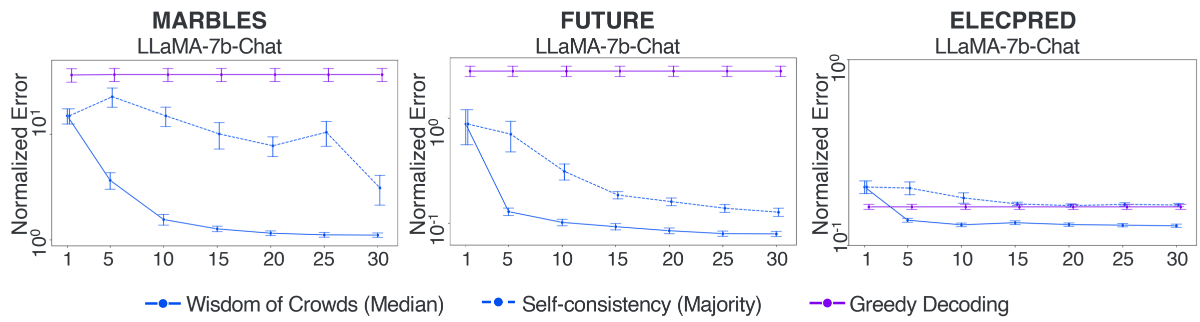

Figure 2: Increased number of sampled reasoning paths boosts wisdom of crowds (median) accuracy, outperforming both self-consistency (majority) and greedy decoding. The trend holds true for all three guesstimation datasets. The normalized error is shown on a logarithmic scale (y axis).The error bars are standard errors calculated based on 30 resampling.

suring cup, a shot glass, a Starbucks iced tall cup, an Altoids tin, and a box for a deck of standard Bicycle playing cards) and three different items (standard-sized U.S. marbles, standard-sized M&Ms, and U.S. quarters). For example, one question asks: *"How many standard-sized U.S. marbles does it take to fill a one-cup dry ingredient measuring cup? Think step-by-step."* The ground-truth answer for each question was determined by manually measuring each quantity three times and taking the median. **Human Experiment.** To replicate previous findings about WOC in human crowds, and compare the LLMs' guesstimation performance with humans, we recruited 230 participants from a U.S. university. Participants received course credit for their participation. Each participant was asked to provide estimates for each question in the MARBLES dataset. We also asked participants to rate their familiarity with each item and container on a 5-point scale (from 1 = "not familiar at all" to 5 = "extremely familiar"). For each question, we only used data from participants who rated their familiarity as at least 4 ("quite familiar") for both the item and the container, yielding an average of 64.9 valid responses per question. We conducted a human experiment only for the MARBLES dataset to ensure genuine guesstimation without easy access to the ground truth, as participants might already know the answers to some questions in the FUTURE and ELECPRED datasets (see below).

FUTURE Dataset The *FUTURE* dataset consists of 15 guesstimation questions about predicting quantities of events in 2024, which was in the future at the time of model training but are now known. These quantities all come from a period after the pretraining cutoff date of the LLMs' training corpora, ensuring that the models could not rely on memorization but instead had to reason based on their world models. For example, one question asks: *"In the second quarter of 2023, the number of vehicles Ford sold was 531,662. In the second quarter of 2024, how many vehicles will Ford sell? Think step-by-step."* The pretraining cutoff dates of all LLMs we considered were before 2024.[1] The true answer for each question was determined based on information from credible websites (§B).

ELECPRED Dataset The *ELECPRED* dataset consists of 51 guesstimation questions, covering 50 U.S. states and Washington, D.C. The task required LLMs to predict the percentage of votes Kamala Harris would receive in the 2024 U.S. presidential election for each state. Similar to the *FUTURE* dataset, the election occurred after all LLMs' pretraining cutoff dates. This ensured that the models could not rely on memorization but instead had to reason based on their world models about factors like demographics, historical trends, and political figures. The ground truth for each state was determined using official election results.

2.2 Large Language Models

We tested the guesstimation capabilities in ten contemporary LLMs, including open-source and proprietary models. We included five LLaMA models (Touvron et al., 2023), a Mistral model (Jiang et al., 2023), two Mixtral models (Jiang et al., 2024), and two GPT models. For the model details, see §A. For our compute resources, see §D.

[1]The only exception was the `Mixtral-8x22b-instruct-v0.1` model, which has a cutoff date in Apr. 2024. Therefore, we excluded it when evaluating it on the FUTURE dataset.

Guesstimation Dataset	Model	Wisdom of Crowds (WOC; Median)	Self-Consistency (Majority)	Greedy
MARBLES	Human Survey	**0.57** [0.54, 0.59]	0.61 [0.57, 0.64]	–
	Mistral			
	mistral-7b-instruct-v0.2	**26.60** [21.39, 31.80]	1154.61 [521.83, 1787.39]	1593.00 [487.33, 2698.67]
	Mixtral			
	mixtral-8x7b-instruct-v0.1	**1.57** [0.84, 2.30]	28.11 [14.35, 41.87]	12.81 [5.05, 20.58]
	mixtral-8x22b-instruct-v0.1	**1.33** [1.13, 1.54]	33.66 [1.78, 65.54]	4.79 [2.24, 7.34]
	LLaMA 2			
	llama-2-7b-chat-hf	**1.25** [0.89, 1.61]	88.44 [1.12, 175.76]	36.80 [7.32, 66.28]
	llama-2-13b-chat-hf	**0.55** [0.47, 0.63]	2.17 [1.17, 3.17]	1.31 [0.84, 1.78]
	llama-2-70b-chat-hf	**0.49** [0.38, 0.61]	1.40 [0.68, 2.11]	29.16 [13.08, 45.24]
	LLaMA 3			
	llama-3.1-8b-instruct	**0.81** [0.76, 0.85]	0.94 [0.91, 0.97]	2.80 [1.75, 3.85]
	llama-3.1-70b-instruct	**0.49** [0.37, 0.61]	1.07 [0.76, 1.39]	6.55 [0.79, 12.30]
	GPT			
	gpt-3.5-turbo-0125	**0.64** [0.53, 0.74]	0.73 [0.50, 0.95]	16.82 [3.72, 29.93]
	gpt-4-0125-preview	**1.00** [0.76, 1.23]	1.07 [0.77, 1.37]	1.04 [0.73, 1.34]
FUTURE	Mistral			
	mistral-7b-instruct-v0.2	**0.61** [0.47, 0.75]	0.91 [0.84, 0.97]	1.79 [0.38, 3.20]
	Mixtral			
	mixtral-8x7b-instruct-v0.1	**0.09** [0.06, 0.12]	**0.09** [0.06, 0.11]	0.60 [0.16, 1.04]
	LLaMA 2			
	llama-2-7b-chat-hf	**0.08** [0.06, 0.11]	1.19 [0.19, 2.18]	2.45 [1.00, 3.89]
	llama-2-13b-chat-hf	**0.09** [0.05, 0.12]	7.53 [1.27, 13.80]	0.11 [0.07, 0.15]
	llama-2-70b-chat-hf	**0.09** [0.06, 0.11]	4.57 [0.41, 8.73]	0.19 [0.11, 0.28]
	LLaMA 3			
	llama-3.1-8b-instruct	**0.54** [0.42, 0.65]	7.84 [1.60, 14.08]	8.54 [2.20, 14.89]
	llama-3.1-70b-instruct	**0.09** [0.06, 0.12]	0.10 [0.07, 0.12]	0.10 [0.07, 0.13]
	GPT			
	gpt-3.5-turbo-0125	**0.10** [0.06, 0.13]	**0.10** [0.06, 0.13]	**0.10** [0.06, 0.13]
	gpt-4-0125-preview	**0.08** [0.06, 0.11]	0.09 [0.07, 0.12]	**0.08** [0.06, 0.11]
ELECPRED	Mistral			
	mistral-7b-instruct-v0.2	**0.07** [0.06, 0.07]	0.11 [0.10, 0.13]	0.16 [0.13, 0.20]
	Mixtral			
	mixtral-8x7b-instruct-v0.1	**0.05** [0.05, 0.06]	0.06 [0.06, 0.07]	0.09 [0.07, 0.11]
	mixtral-8x22b-instruct-v0.1	**0.06** [0.05, 0.07]	**0.06** [0.06, 0.07]	0.12 [0.10, 0.13]
	LLaMA 2			
	llama-2-7b-chat-hf	**0.14** [0.12, 0.16]	0.16 [0.15, 0.18]	0.16 [0.13, 0.19]
	llama-2-13b-chat-hf	**0.10** [0.09, 0.11]	0.12 [0.11, 0.13]	0.16 [0.12, 0.19]
	llama-2-70b-chat-hf	**0.11** [0.09, 0.12]	0.12 [0.11, 0.14]	0.12 [0.11, 0.13]
	LLaMA 3			
	llama-3.1-8b-instruct	**0.07** [0.06, 0.07]	0.08 [0.07, 0.09]	0.08 [0.07, 0.08]
	llama-3.1-70b-instruct	**0.05** [0.05, 0.06]	**0.05** [0.05, 0.06]	0.08 [0.06, 0.10]
	GPT			
	gpt-3.5-turbo-0125	**0.07** [0.06, 0.07]	0.08 [0.07, 0.08]	0.16 [0.12, 0.20]
	gpt-4-0125-preview	**0.05** [0.05, 0.06]	**0.05** [0.04, 0.05]	**0.05** [0.05, 0.06]

Table 1: Normalized errors (ε) for each model on three guesstimation tasks MARBLES, FUTURE, and ELECPRED. The table is organized by model families and shows results under three decoding strategies: Wisdom of Crowds (WOC; median), Self-Consistency (majority), and Greedy decoding. Brackets denote standard errors. Notably, WOC is consistently the best decoding method.

2.3 Decoding Methods for Guesstimation

For each guesstimation question, an LLM generates a response $x \in \mathbb{N}$, where there exists a ground truth $x^* \in \mathbb{N}$. We evaluate three decoding methods for LLM's guesstimation: *wisdom of crowds* (WOC) decoding, *self-consistency* decoding, and *greedy decoding* decoding. For the WOC and self-consistency methods, given a question, we sample n reasoning paths (using chain-of-thought prompting; Wei et al., 2022b,a) from the LLM using temperature sampling with $T = 1$ (Figure 1). Each reasoning path yields a corresponding estimate x, resulting in a set of responses denoted as $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. For WOC, we take the median of the response set, $\text{median}(\mathcal{X}) = x_{\lceil \frac{n}{2} \rceil}$, as the final estimate. For self-consistency, we calculate the mode of the response set, $\text{mode}(\mathcal{X})$. In cases where the response set has multiple modes, we randomly choose one. For greedy decoding, the temperature is set to 0, making the response deterministic. Thus, for each question, we obtain only one response from an LLM.

2.4 Evaluation Metric

To assess the accuracy of the estimates across questions, we defined the normalized error. Formally, for a given estimate $\hat{x}$ and its corresponding ground truth x^*, the normalized error ε is defined

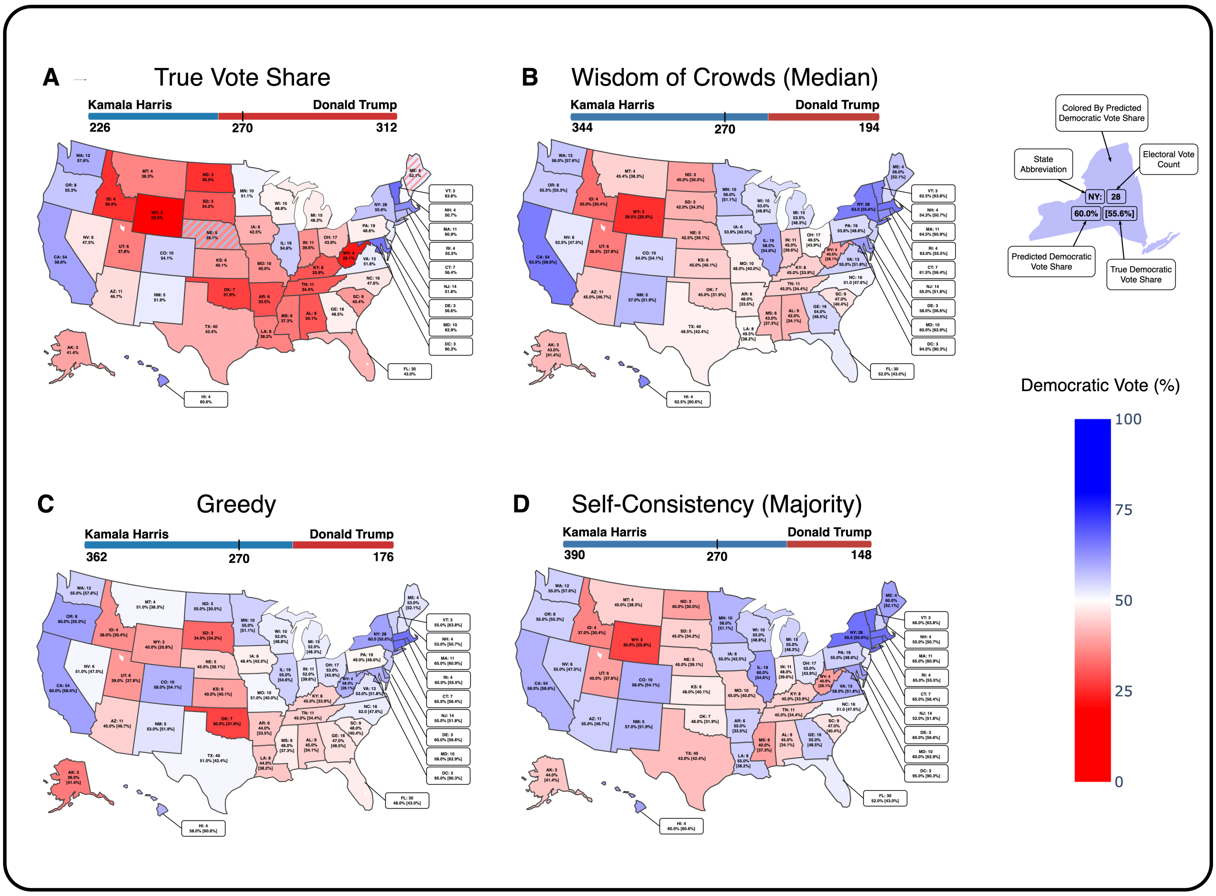

Figure 3: Comparison of Actual and Predicted Vote Percentages in the 2024 U.S. Presidential Election (LLaMA-2-7b-Chat; ELECPRED dataset). (A) The actual vote percentage Kamala Harris received in each state in 2024 US presidential election. (B) The predicted vote percentage using wisdom of crowds (median) decoding. (C) The predicted vote percentage using greedy decoding. (D) The predicted vote percentage using self-consistency (majority) decoding. For (B), (C), and (D) the predicted vote percentage using each strategy is given, followed by the actual vote percentage in brackets.

as: $\varepsilon = |\hat{x} - x^*|/x^*$. This metric is commonly used in previous literature on guesstimation tasks in human studies (Becker et al., 2017, 2019).

3 Results

Humans are Good at Guesstimation Human crowds achieve highly accurate guesstimation under WOC decoding ($\varepsilon = 0.57$) compared to most LLMs in the MARBLES dataset (Table 1). This replicates previous findings about WOC in humans (Galton, 1907; Yu et al., 2018). In addition, WOC decoding has a higher accuracy compared to self-consistency decoding ($\varepsilon : 0.57 < 0.61$).

Wisdom of Crowds (WOC) Decoding Supports Guesstimation in LLMs For LLMs, the WOC decoding method consistently outperforms the self-consistency and greedy decoding methods in the three guesstimation tasks and across different model variants (Table 1). In a few cases, self-consistency and/or greedy decoding achieves the same accuracy as WOC decoding, but WOC is consistently among the best decoding methods.

WOC Performance Improves More Efficiently Than Self-Consistency with More Sampled Reasoning Paths Increasing the number of sampled reasoning paths consistently improves the accuracy of the WOC decoding method (Figure 2). In contrast, while increasing the sample size also leads to better guesstimation performance of the self-consistency method, the improvement is much slower and less efficient than the WOC decoding method. For example, for both FUTURE and ELECPRED datasets, WOC decoding using 5 samples achieves higher accuracy than self-consistency decoding using 30 samples.

WOC Decoding Produces the Most Accurate Prediction of the 2024 U.S. Presidential Election As shown in Table 1, WOC decoding outperforms both self-consistency and greedy decoding in prediction accuracy in terms of the vote percentage Kamala Harris received in the 2024 U.S. presidential election. However, the difference in quality is difficult to interpret intuitively. To better illustrate the results, we visualized the predicted election outcomes on a national map (Figure 3). While LLMs predicted the percentage of votes Kamala Harris would receive in each state, we converted these percentages into electoral votes to compare them with the actual election outcome, in which Donald Trump won 312 electoral votes, while Kamala Harris received 226. The results show that WOC decoding provided the closest prediction (194 electoral votes for Harris). In contrast, greedy decoding predicted 176, and self-consistency predicted 148. Notably, both greedy decoding and self-consistency made implausible errors: greedy decoding predicted a Democratic win in Texas, and self-consistency incorrectly predicted Democratic wins in Arkansas and Louisiana. While WOC decoding achieved the most accurate prediction, it showed an overall bias favoring Democrats. Understanding the source of this bias remains an open question for future research.

4 Related Work

Guesstimation and Wisdom of Crowds For a crowd to reach better guesstimation, wisdom of crowds (WOC) has proven to be effective, as long as individual estimates within these groups are statistically independent (Surowiecki, 2005; Nofer and Nofer, 2015). This independence ensures that their errors are uncorrelated, allowing them to cancel out in aggregate. WOC has shown applications in real-world guesstimation challenges like market prediction and political forecasting (Yu et al., 2018).

Prompting and Decoding Strategies for LLM Reasoning Prompting-based methods are designed to guide large language models (LLMs) in generating desired outputs. Wei et al. (2022b) introduced chain-of-thought (CoT) prompting to explicitly elicit intermediate reasoning steps, which significantly improves performance on tasks requiring reasoning. Kojima et al. (2022) further extended CoT prompting to zero-shot settings. However, the variability in generated chain-of-though reasoning responses has motivated researchers to explore more robust decoding strategies. For example, Wang et al. (2023) proposed the "self-consistency" decoding approach that samples multiple reasoning paths and selects the most consistent answer, leading to better quality and accuracy than greedy decoding. However, subsequent work showed that the self-consistency is not always effective (Nguyen et al., 2024; Byerly and Khashabi, 2024). To our best understanding, we are the first to apply WOC decoding strategy to LLM reasoning responses.

5 Conclusion

In this study, we show that LLMs possess a world model necessary for effective guesstimation, a common yet overlooked task in the AI community. To evaluate this, we introduce three guesstimation datasets: *MARBLES*, *FUTURE*, and *ELECPRED*, where one must estimate both concrete and abstract quantities based on knowledge about the world. Similar to humans, LLMs also exhibit the WOC effect, in which the median of estimates leads to more accurate results than greedy decoding and self-consistency. In addition, WOC performance improves more efficiently than self-consistency as the number of sampled reasoning paths increases. In sum, we introduce guesstimation as a new task that is very common in the real world yet has been largely overlooked by the AI community.

Limitations

The Scope of Guesstimation Questions is U.S.-Centric Our guesstimation questions are heavily U.S.-centric, covering topics such as common U.S. household items, U.S. economic statistics, and U.S. election results. It remains unclear whether LLMs would perform equally well on guesstimation tasks in other cultural and geographical contexts. Future work should explore the generalizability of these findings across different cultural contexts.

Mechanism Behind WOC's Superiority While we find that WOC decoding consistently outperforms self-consistency, the underlying mechanism driving this improvement remains unclear. One possible explanation is that taking the median helps mitigate the influence of extreme outlier predictions, making WOC more robust. However, a deeper investigation is needed to fully understand why WOC is superior and whether similar effects hold across different types of reasoning tasks.

Ethics Statement

For the human experiment, our study has been reviewed and approved by the Institutional Review Board (IRB) of our institution. In addition, we will release our code base solely for research purposes, and adhere to the terms of use by OpenAI's API [2] and their MIT license [3], as well as Mistral AI's non-production license (MNPL) [4] and Meta's Llama community license [5].

[2] https://openai.com/policies/terms-of-use

[3] https://github.com/openai/openai-openapi/blob/master/LICENSE

[4] https://mistral.ai/licenses/MNPL-0.1.md

[5] https://www.llama.com/faq/

References

Joshua Becker, Devon Brackbill, and Damon Centola. 2017. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 114(26):E5070.

Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 116(22):10717–10722.

Adam Byerly and Daniel Khashabi. 2024. How effective is self-consistency for long-context problems? *arXiv preprint arXiv:2411.01101*.

Francis Galton. 1907. Vox populi. *Nature*, 75(1949):450–451.

History, Art & Archives, U.S. House of Representatives. Election Statistics: 1920 to Present. Accessed: February 11, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Takeru Kojima, Shixiang Gu, Mike Reid, Yutaka Matsuo, and Kazuto Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Alex Nguyen, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2024. When is the consistent prediction likely to be a correct prediction? *arXiv preprint arXiv:2407.05778*.

Michael Nofer and Michael Nofer. 2015. Are crowds on the internet wiser than experts?–the case of a stock prediction community. *The Value of Social Media for Predicting Stock Returns: Preconditions, Instruments and Performance Analysis*, pages 27–61.

James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chao Yu, Yueting Chai, and Yi Liu. 2018. Literature review on collective intelligence: a crowd science perspective. *International Journal of Crowd Science*, 2(1):64–73.

A Selection of the LLMs

Table 2 lists the LLMs that we evaluate. The knowledge cutoff dates were decided based on the model description webpage. For the Mistral and Mixtral models, the knowledge cutoff dates were not released, so the date listed is the date of model weight commits on HuggingFace [6][7][8].

Model Family	Model Variant	Knowledge Cutoff Date
Mistral	mistral-7b-instruct-v0.2	Before Dec. 2023
Mixtral	mixtral-8x7b-instruct-v0.1	Before Dec. 2023
	mixtral-8x22b-instruct-v0.1	Before Apr. 2024
LLaMA 2	llama-2-7b-chat-hf	Jul. 2023
	llama-2-13b-chat-hf	Jul. 2023
	llama-2-70b-chat-hf	Jul. 2023
LLaMA 3.1	llama-3.1-8b-instruct	Dec. 2023
	llama-3.1-70b-instruct	Dec. 2023
GPT	gpt-3.5-turbo-0125	Sep. 2021
	gpt-4-0125-preview	Dec. 2023

Table 2: List of large language models.

B Guesstimation Questions and Ground Truth Answers

Tables 3 and 4 list the guesstimation questions used in the MARBLES and FUTURE datasets along with their corresponding ground truth answers.

The following sources were used to determine the ground truth answers for the FUTURE dataset:

- Ford Sales
- New York City Population
- 2024 Olympic Medal Table, 2020 Olympic Medal Table
- United States GDP
- Tesla Sales
- University of Wisconsin-Madison Enrollment
- Apple 2024 Sales, Apple 2023 Sales
- New Jersey 2024 Temperature, New Jersey 2023 Temperature

[6]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2/commit/dca6e4b60aca009ed25ffa70c9bb65e46960a573

[7]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1/commit/858fdc292793fc3e671bf51fc5586c5cc10fbe3a

[8]https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1/commit/796bc4393fd5e7e0c0ff1c44de2526419f163003

- Sony Sales
- 2023 Forest Loss, 2022 Forest Loss
- 2023 Satellite Launches, 2024 Satellite Launches
- United States Home Prices
- United States Unemployment Claims
- 2024 TSA Passenger Count, 2023 TSA Passenger Count

Table 5 lists the percentage of the vote Kamala Harris received in the 2024 presidential Election and number of electoral votes for each state and the District of Columbia.

The following is text is the format of the prompt for the ELECPRED dataset, where the results are listed for all presidential elections from 1976 to 2020:

```
Here is a history of prior voting results from the US state of Alabama for US Presidential elections:

1976: Jimmy Carter (Democrat) versus Gerald Ford (Republican). Carter (the Democrat) received 56 percent of the vote.

...

2020: Joseph R. Jr Biden (Democrat) versus Donald J. Trump (Republican). Biden (the Democrat) received 37 percent of the vote.

In the 2024 election, the candidates will be Vice President Kamala Harris (the Democrat) and former President Donald Trump (the Republican). What percentage of the vote in Alabama do you think Kamala Harris (the Democrat) will receive? You must not predict a tie.
```

The historical results from each state can be found on the United States House of Representatives Archive (History, Art & Archives, U.S. House of Representatives).

C The Prompts used for querying the LLMs

Table 6 lists the prompts that are used when querying the LLMs on the MARBLES dataset. Table 8 lists the prompts that are used when querying the LLMs on the ELECPRED dataset. Table 7 lists the prompts that are used when querying the LLMs on the FUTURE dataset. Note the addition of the phrase "If you don't have enough information, just make a guess." to the FUTURE system prompts.

D Compute Resources

We ran all experiments on a GPU machine equipped with 2x NVIDIA A100.

Question	True Answer
How many standard-sized U.S. marbles does it take to fill a one cup dry ingredient measuring cup?	62
How many standard-sized U.S. marbles does it take to fill a single-shot shot glass?	13
How many standard-sized U.S. marbles does it take to fill a Starbucks iced tall cup?	109
How many standard-sized U.S. marbles does it take to fill an Altoids tin container?	22
How many standard-sized U.S. marbles does it take to fill the box for a deck of cards (standard-sized Bicycle playing cards)?	24
How many standard-sized M&Ms does it take to fill a one cup dry ingredient measuring cup?	210
How many standard-sized M&Ms does it take to fill a single-shot shot glass?	51
How many standard-sized M&Ms does it take to fill a Starbucks iced tall cup?	382
How many standard-sized M&Ms does it take to fill an Altoids tin container?	95
How many standard-sized M&Ms does it take to fill the box for a deck of cards (standard-sized Bicycle playing cards)?	96
How many U.S. quarters does it take to fill a one cup dry ingredient measuring cup?	160
How many U.S. quarters does it take to fill a single-shot shot glass?	42
How many U.S. quarters does it take to fill a Starbucks iced tall cup?	280
How many U.S. quarters does it take to fill an Altoids tin container?	70
How many U.S. quarters does it take to fill the box for a deck of cards (standard-sized Bicycle playing cards)?	70

Table 3: List of all MARBLES questions and their corresponding true answers.

Question	True Answer
In the second quarter of 2023, the number of vehicles Ford sold was 531662. In the second quarter of 2024, how many vehicles will Ford sell?	536,050
In 2023 the population of the New York City Metropolitan Area was 18937000. In 2024, how many people will live in the New York City Metropolitan Area?	19,034,000
In the 2020 Summer Olympics, the number of medals the United States won was 113. In the 2024 Summer Olympics, how many medals will the United States win?	126
In Q2 2023, the United States' GDP in billions was 27453.815. In Q2 2024, how many billions will the United States' GDP be?	29,016.714
In Q1 2023, Tesla's total revenue in billions was 23.329. In Q1 2024, how many billions will Tesla's total revenue be?	21.301
In the 2023-24 school year, the number of students enrolled at the University of Wisconsin Madison was 50,633. In the 2024-25 school year, how many students will be enrolled at the University of Wisconsin Madison?	52,097
In Q1 2023 Apple's total revenue in billions 117.2. In Q1 2024, how many billions will Apple's total revenue be?	119.6
The average temperature in degrees Fahrenheit in New Jersey in June 2023 was 67.8. In June 2024, what will the average temperature in degrees Fahrenheit in New Jersey be?	73.6
In Q1 2023 the number of PlayStation 5 units sold was 3300000. In Q1 2024, how many PlayStation 5 units will be sold?	2,400,000
In Q1 2023 the number of monthly active users on the PlayStation Network in millions was 108. In Q1 2024, how many monthly active users in millions will the PlayStation Network have?	116
In 2022 the number of acres of primary tropical forest lost was 10130000. In 2023, how many acres of primary tropical forest will be lost?	9,100,000
The number of satellites the United States launched into space from January to October 2023 was 85. From January to October 2024, how many satellites will the United States launch into space?	111
In Q1 2023 the average sale price of a house in the United States was 505300. In Q1 2024, what will the average sale price of a house in the United States be?	519,700
In Q3 2023 the number of unemployment insurance claims filed was 232643. In Q3 2024, how many unemployment insurance claims will be filed?	231,154
From January 2023 to the beginning of October 2023 the number of passengers that passed through TSA security in the United States was 638549095. From January 2024 to the beginning of October 2024, how many passengers will pass through TSA security in the United States?	677,657,486

Table 4: List of all FUTURE questions and their corresponding true answers.

State	Electoral Vote Count	% Harris Vote
Alabama	9	34.1%
Alaska	3	41.4%
Arizona	11	46.7%
Arkansas	6	33.5%
California	54	58.6%
Colorado	10	54.1%
Connecticut	7	56.4%
Delaware	3	56.6%
District Of Columbia	3	90.3%
Florida	30	43.0%
Georgia	16	48.5%
Hawaii	4	60.6%
Idaho	4	30.4%
Illinois	19	54.6%
Indiana	11	39.6%
Iowa	6	42.5%
Kansas	6	41.0%
Kentucky	8	33.9%
Louisiana	8	38.2%
Maine	4	52.1%
Maryland	10	62.9%
Massachusetts	11	60.9%
Michigan	15	48.3%
Minnesota	10	51.1%
Mississippi	6	37.3%
Missouri	10	40.0%
Montana	4	38.3%
Nebraska	5	39.1%
Nevada	6	47.5%
New Hampshire	4	50.7%
New Jersey	14	51.8%
New Mexico	5	51.9%
New York	28	55.6%
North Carolina	16	47.6%
North Dakota	3	30.5%
Ohio	17	43.9%
Oklahoma	7	31.9%
Oregon	8	55.3%
Pennsylvania	19	48.6%
Rhode Island	4	55.5%
South Carolina	9	40.4%
South Dakota	3	34.2%
Tennessee	11	34.4%
Texas	40	42.4%
Utah	6	37.8%
Vermont	3	63.8%
Virginia	13	51.8%
Washington	12	57.6%
West Virginia	4	28.1%
Wisconsin	10	48.8%
Wyoming	3	25.8%

Table 5: List of all state results for the United States 2024 presidential election.

Prompt Type	Message Type	Prompt	Example
Initial Prompt	*System Message*	You must provide a final answer.	You must provide a final answer.
Initial Prompt	*User Message*	{question} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]	{How many standard-sized M&Ms does it take to fill a Starbucks iced tall cup?} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]
Two Step Extraction	*User Message*	{initial_response}. Therefore the final answer (arabic numerals) is	{How many standard-sized M&Ms does it take to fill a Starbucks iced tall cup? Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation] Reasoning: A Starbucks iced tall cup has a volume of approximately 12 oz or 355 ml. The volume of a single standard-sized M&M is estimated to be around 0.103 oz or 2.94 ml based on the density of milk chocolate and average dimensions of the candy. To calculate the number of M&Ms needed to fill the cup, we can convert the total volume to M&M volumes and round up to the nearest M&M to account for excess candy: Number of M&Ms = Total volume / Volume of a single M&M Number of M&Ms = 355 ml / 2.94 ml Number of M&Ms = 121.63 = 122 M&Ms Final answer: 122 M&Ms.} Therefore the final answer (arabic numerals) is

Table 6: The prompts used for query the LLMs on the MARBLES dataset.

Prompt Type	Message Type	Prompt	Example
Initial Prompt	*System Message*	You must provide a final answer. If you don't have enough information, just make a guess.	You must provide a final answer. If you don't have enough information, just make a guess.
Initial Prompt	*User Message*	{question} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]	{In the second quarter of 2023, the number of vehicles Ford sold was 531662. In the second quarter of 2024, how many vehicles will Ford sell?} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]
Two Step Extraction	*User Message*	{initial_response}. Therefore the final answer (arabic numerals) is	{In the second quarter of 2023, the number of vehicles Ford sold was 531662. In the second quarter of 2024, how many vehicles will Ford sell? Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation] Answer : 564250 Reasoning : The information given in the question is Second quarter of 2023 - Ford sold 531662.} Therefore the final answer (arabic numerals) is

Table 7: The prompts used for query the LLMs on the FUTURE dataset.

Prompt Type	Message Type	Prompt	Example
Initial Prompt	*System Message*	You must provide a final answer.	You must provide a final answer.
Initial Prompt	*User Message*	{question} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]	{Here is a history of prior voting results from the US state of Alabama for US Presidential elections: 1976: Jimmy Carter (Democrat) versus Gerald Ford (Republican). Carter (the Democrat) received 56 percent of the vote. 1980: Jimmy Carter (Democrat) versus Ronald Reagan (Republican). Carter (the Democrat) received 49 percent of the vote. 1984: Walter Mondale (Democrat) versus Ronald Reagan (Republican). Mondale (the Democrat) received 38 percent of the vote. 1988: Michael Dukakis (Democrat) versus George H.W. Bush (Republican). Dukakis (the Democrat) received 40 percent of the vote. 1992: Bill Clinton (Democrat) versus George H.W. Bush (Republican). Clinton (the Democrat) received 46 percent of the vote. 1996: Bill Clinton (Democrat) versus Robert Dole (Republican). Clinton (the Democrat) received 46 percent of the vote. 2000: Al Gore (Democrat) versus George W. Bush (Republican). Gore (the Democrat) received 42 percent of the vote. 2004: John Kerry (Democrat) versus George W. Bush (Republican). Kerry (the Democrat) received 37 percent of the vote. 2008: Barack H. Obama (Democrat) versus John Mccain (Republican). Obama (the Democrat) received 39 percent of the vote. 2012: Barack H. Obama (Democrat) versus Mitt Romney (Republican). Obama (the Democrat) received 38 percent of the vote. 2016: Hillary Clinton (Democrat) versus Donald J. Trump (Republican). Clinton (the Democrat) received 35 percent of the vote. 2020: Joseph R. Jr Biden (Democrat) versus Donald J. Trump (Republican). Biden (the Democrat) received 37 percent of the vote. In the 2024 election, the candidates will be Vice President Kamala Harris (the Democrat) and former President Donald Trump (the Republican). What percentage of the vote in Alabama do you think Kamala Harris (the Democrat) will receive? You must not predict a tie.} Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation]
Two Step Extraction	*User Message*	{initial_response}. Therefore the final answer (arabic numerals) is	{Here is a history of prior voting results from the US state of Alabama for US Presidential elections: 1976: Jimmy Carter (Democrat) versus Gerald Ford (Republican). Carter (the Democrat) received 56 percent of the vote. 1980: Jimmy Carter (Democrat) versus Ronald Reagan (Republican). Carter (the Democrat) received 49 percent of the vote. 1984: Walter Mondale (Democrat) versus Ronald Reagan (Republican). Mondale (the Democrat) received 38 percent of the vote. 1988: Michael Dukakis (Democrat) versus George H.W. Bush (Republican). Dukakis (the Democrat) received 40 percent of the vote. 1992: Bill Clinton (Democrat) versus George H.W. Bush (Republican). Clinton (the Democrat) received 46 percent of the vote. 1996: Bill Clinton (Democrat) versus Robert Dole (Republican). Clinton (the Democrat) received 46 percent of the vote. 2000: Al Gore (Democrat) versus George W. Bush (Republican). Gore (the Democrat) received 42 percent of the vote. 2004: John Kerry (Democrat) versus George W. Bush (Republican). Kerry (the Democrat) received 37 percent of the vote. 2008: Barack H. Obama (Democrat) versus John Mccain (Republican). Obama (the Democrat) received 39 percent of the vote. 2012: Barack H. Obama (Democrat) versus Mitt Romney (Republican). Obama (the Democrat) received 38 percent of the vote. 2016: Hillary Clinton (Democrat) versus Donald J. Trump (Republican). Clinton (the Democrat) received 35 percent of the vote. 2020: Joseph R. Jr Biden (Democrat) versus Donald J. Trump (Republican). Biden (the Democrat) received 37 percent of the vote. In the 2024 election, the candidates will be Vice President Kamala Harris (the Democrat) and former President Donald Trump (the Republican). What percentage of the vote in Alabama do you think Kamala Harris (the Democrat) will receive? You must not predict a tie. Think step-by-step. You have to use the following format Reasoning: [Your step-by-step reasoning] Final answer: [A number. No other text or explanation] Reasoning: Alabama has consistently voted for the Republican candidate in US Presidential elections. The state has a voter population of 3,894,973.} Therefore the final answer (arabic numerals) is

Table 8: The prompts used for query the LLMs on the ELECPRED dataset.