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

Yun-Shiuan Chuang **Nikunj Harlalka**[†] **Sameer Narendran**[†] **Alexander Cheung**
Sizhe Gao **Siddharth Suresh** **Junjie Hu** **Timothy T. Rogers**
University of Wisconsin-Madison
{yunshiuan.chuang, nirunwiroj, studdiford, agoyal25}@wisc.edu
{vfrigo, syang84, dshah, junjie.hu, ttrogers}@wisc.edu

Abstract

Guesstimation, the task of making approximate quantity estimates, is a common real-world challenge. However, it has been largely overlooked in large language models (LLMs) and vision language models (VLMs) research. We introduce a novel guesstimation dataset, *MARBLES*. This dataset requires one to estimate how many items (e.g., marbles) can fit into containers (e.g., a one-cup measuring cup), both with and without accompanying images. 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 "WOC decoding" strategy for LLM guesstimation. We show that LLMs/VLMs perform well on guesstimation, suggesting that they possess some level of a "world model" necessary for guesstimation. Moreover, similar to human performance, the WOC decoding method improves LLM/VLM guesstimation accuracy. Furthermore, the inclusion of images in the multimodal condition enhances model performance. These results highlight the value of WOC decoding strategy for LLMs/VLMs and position guesstimation as a probe for evaluating LLMs/VLMs' world model.

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 [3, 17]. WOC phenomena are thought to rely on the grounding of conceptual knowledge in embodied, multi-modal experience. For instance, when estimating the number of jelly-beans in a jar [10], 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–properties experienced directly through perception and action in the world, in addition to being expressed in language.

Here we assess whether contemporary large language models (LLMs) exhibit WOC phenomena similar to those observed in human populations. On one hand, 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. On the other hand, LLMs do not have the same world-grounded experience thought to inform human WOC phenomena: language-only models are trained solely on tokens generated from text, while vision language models (VLMs) additionally

[†] Joint second authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

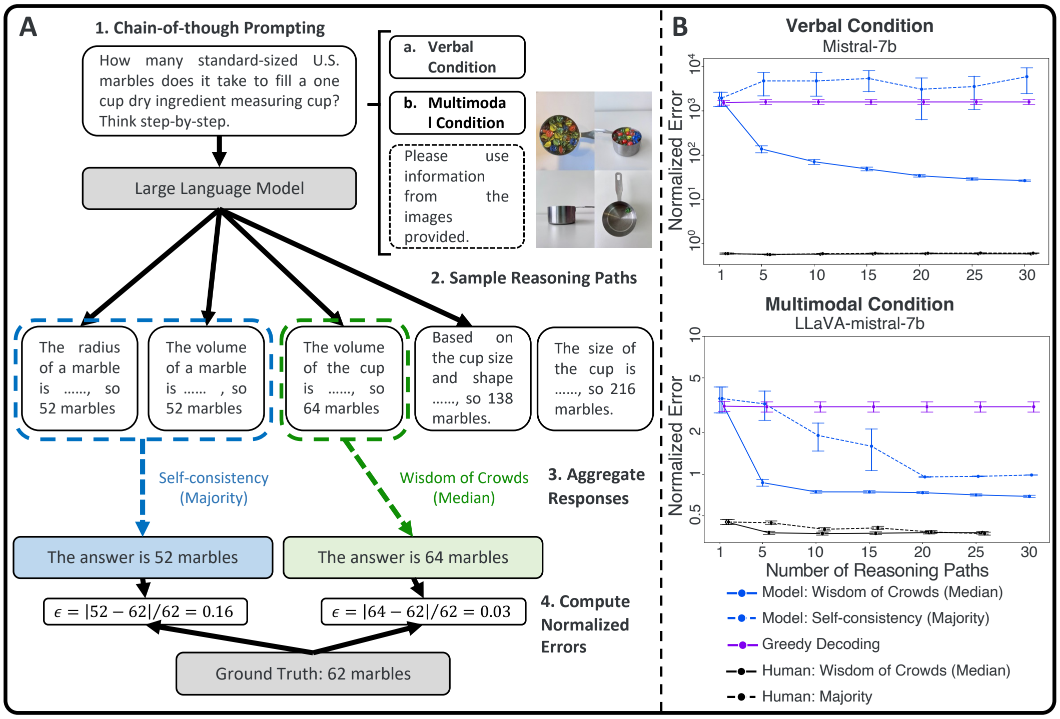

Figure 1: (A) The steps of LLM/VLM guesstimation through self-consistency decoding method and wisdom of crowd (WOC) decoding method, across both verbal condition and multimodal condition. (B) Increased number of sampled reasoning paths boosts wisdom of crowds (median) accuracy, outperforming both self-consistency (majority) and greedy decoding. The trend holds true across (top) verbal condition and (bottom) multimodal condition. The normalized error is shown on a logarithmic scale (y axis).The error bars are standard errors calculated based on 30 resampling.

incorporate static images. If WOC phenomena arise from the embodiment of concepts in aspects of physical experience–the ability to perceive and act on the world, interact with objects, move through space, and so on–one might expect LLMs and VLMs to exhibit patterns of behavior quite different from crowds of humans. We therefore conducted a series of estimation experiments with human participants and a range of LLM/VLM models, first establishing the key phenomena for people, then assessing whether similar or different patterns arise in large language models.

In all experiments, the agent (human or model) was asked to estimate the largest quantity of items that could fit in a specified container. The items and the containers were all familiar, rigid everyday objects with a standard shape and size that could be referred to / described in American English. For instance, participants might be asked how many standard US marbles can fit in a one-cup dry-ingredients measuring cup, or how many pennies could fit in a one-shot shot glass. We conducted a *verbal* condition, where all instructions were provided in natural language alone, and a *multimodal* condition where the same instructions were accompanied by photographs of the scenario (e.g. picture of a measuring-cup filled with marbles). To quantify the WOC effect in each case, we took the normalized error: absolute difference between median guess and ground truth divided by the ground truth. The more this error terms reduced with increasing size of the crowds, the greater the WOC advantage relative to an individual guesser.

For the human data, we considered how the WOC error varies with larger numbers of participants in the crowd, for both language-only and language+vision variants of the experiment. For LLM/VLM data we conducted parallel analyses, measuring WOC error for increasing samples of the responses generated by a common prompt. We further compared LLM WOC behavior to an alternative *self-consistency* strategy for improving LLM model alignment, which samples model behavior many times and returns the majority-vote across samples, rather than the median. Prior work has suggested that self-consistency can improve model reasoning behavior [12].

2 Methods and Experimental Setup

MARBLES Dataset. Our *MARBLES* dataset consists of 15 guesstimation questions, involving five different containers (a one-cup dry ingredient measuring 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, *"How many standard-sized U.S. marbles does it take to fill a one-cup dry ingredient measuring cup?"*. The true answer for each question was determined by manually measuring the quantity three times and taking the median. Additionally we captured four photographs for each question: a top view with the items filling the container, a tilted view, a side view of the container, and a photo showing a single item inside the container (Figure 1). In the multimodal condition, these images were presented alongside the textual questions, while in the verbal condition, only the textual questions were provided. See §A for the full list of questions and §B for the prompts.

Human Experiment. We recruited 230 participants from a university in the US. Participants were randomly assigned to either the verbal condition (112 participants) or the multimodal condition (108 participants). Each participant was asked to generate 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. This results in 64.9 valid response on average per question.

Large Language Models. We tested various LLMs, including both open-source and proprietary models. For the verbal condition, we included a Mistral model [4], two Mixtral models [5], five LLaMA models [11], three Vicuna models [19], and two GPT models. For the multimodal condition, we included three LLaVA models [7, 6], each with a corresponding base model considered in the verbal condition. See §C for the detailed names of the LLMs.

LLM Guestimation Methods. 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 guesstimation methods for LLM's responses: *wisdom of crowds* (WOC), *self-consistency*, and *greedy decoding*. For the WOC and self-consistency methods, given a question, we sample n reasoning paths (using chain-

Table 1: Normalized errors (ε) averaged across questions for both verbal and multimodal conditions on the guesstimation task (MARBLES). The three columns are the three guesstimation methods. Brackets denote standard errors.

Condition	Model	Wisdom of Crowds (WOC; Median)	Self-Consistency (Majority)	Greedy
Verbal	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]	157.94 [102.55, 213.34]	1593.00 [487.33, 2698.67]
	Mixtral			
	mixtral-8x7b-instruct-v0.1	**1.57** [0.84, 2.30]	5.63 [3.02, 8.24]	12.81 [5.05, 20.58]
	mixtral-8x22b-instruct-v0.1	**1.33** [1.13, 1.54]	1.76 [1.40, 2.13]	4.79 [2.24, 7.34]
	LLaMA 2			
	llama-2-7b-chat-hf	**1.22** [0.89, 1.56]	1.59 [0.77, 2.40]	34.42 [6.86, 61.97]
	llama-2-13b-chat-hf	**0.54** [0.46, 0.62]	1.70 [1.08, 2.32]	1.27 [0.87, 1.67]
	llama-2-70b-chat-hf	**0.49** [0.38, 0.61]	0.76 [0.64, 0.89]	29.16 [13.08, 45.24]
	LLaMA 3			
	llama-3.1-8b	**0.72** [0.66, 0.79]	0.92 [0.89, 0.96]	inf [nan, nan]
	llama-3.1-70b	**0.79** [0.74, 0.83]	0.90 [0.87, 0.94]	inf [nan, nan]
	Vicuna			
	vicuna-7b-v1.5	**0.75** [0.67, 0.84]	0.80 [0.74, 0.87]	8.43 [2.96, 13.90]
	vicuna-13b-v1.5	**0.95** [0.71, 1.19]	1.00 [0.78, 1.23]	4.78 [1.97, 7.59]
	vicuna-33b-v1.3	**0.96** [0.74, 1.18]	1.03 [0.87, 1.20]	7.43 [1.43, 13.42]
	GPT			
	gpt-3.5-turbo-0125	**0.64** [0.53, 0.74]	0.75 [0.50, 1.00]	16.82 [3.72, 29.93]
	gpt-4-0125-preview	**1.00** [0.76, 1.23]	**1.00** [0.69, 1.30]	1.04 [0.73, 1.34]
Multimodal	Human Survey	**0.33** [0.29, 0.38]	**0.33** [0.28, 0.38]	–
	LLaVA			
	llava-v1.6-mistral-7b	**0.69** [0.61, 0.77]	0.98 [0.96, 1.00]	3.09 [1.60, 4.57]
	llava-v1.6-vicuna-7b	**0.75** [0.65, 0.85]	1.27 [0.77, 1.77]	13.10 [2.09, 24.11]
	llava-v1.6-vicuna-13b	**0.66** [0.59, 0.73]	0.85 [0.80, 0.90]	31.60 [10.65, 52.55]

of-thought prompting [14, 13]) 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, median$(\mathcal{X}) = x_{\lceil \frac{n}{2} \rceil}$, as the final estimate. For self-consistency, we calculate the mode of the response set, 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.

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 as: $\varepsilon = |\hat{x} - x^*|/x^*$.

3 Results

Humans are Good at Guesstimation. Across verbal and multimodal conditions, humans achieve the most accurate guesstimation compared to almost all LLMs/VLMs (Table 1). Moreover, human's accuracy is further improved with the presence of images in the multimodal condition ($\varepsilon = 0.33$). In addition, in multimodal condition, the error ε of WOC decoding reduces with increasing size of the crowds (Figure 1; ε reduces from 0.45 to 0.33 when crowd size increases from 1 to 25). Interesting, such WOC reduction does not hold true for verbal condition (ε remains 0.59).

Wisdom of Crowds (WOC) Decoding Supports Guesstimation in LLMs/VLMs. For LLMs/VLMs, the WOC decoding method consistently outperforms the self-consistency and greedy decoding methods across model sizes and variants (Table 1). The Mistral and Mixtral models enjoy the largest gain with WOC. The only exception is `gpt-4-0125-preview`, where WOC and self-consistency has the same performance.

Increasing Number of Sampled Reasoning Paths Enhances Wisdom of Crowds Performance. Increasing the number of sampled reasoning paths consistently improves the accuracy of the WOC method (Figure 1) across both the verbal and multimodal conditions. In contrast, increasing the sample size does not consistently lead to better guesstimation performance of self-consistency method.

Multimodal Inputs Improve Guesstimation Performance. Similar to human, LLMs also perform better in the multimodal condition, where both text and images are provided as input. For instance, as shown in Table 1, the LLaVA model, which takes as input as both text and images and is powered by the Mistral-7B base model, significantly outperforms its text-only counterpart, Mistral-7B-Instruct. This highlights that multimodal information improve the LLM's world model.

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 [10, 8]. 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 [17].

Vision language models (VLMs)' Spatial Reasoning. Previous work has investigated the spatial reasoning capabilities of vision language models (VLMs). Explicit grounding of the model with spatial awareness helps the model perform better in spatial reasoning [1, 18, 9, 15, 2, 16]. However, to our knowledge, no work to date has investigated VLMs' capabilities in guesstimation.

5 Conclusion

In the study, we show that LLMs/VLMs possess the world model necessary for effective guesstimation, a common yet overlooked task in the AI community. To evaluate this, we introduce the *MARBLES* dataset, where one needs to estimate how many items can fit into various containers (along side with photos included for the multimodal condition). We show that humans are good at guesstimation, and their accuracy is further improved by the inclusion of images. Moreover, human WOC effect emerges in the multimodal condition. Second, similar to human, LLMs/VLMs also show the

WOC effect, where the median of estimates leads to more accurate results than greedy decoding or self-consistency. In addition, like human, including visual information further improved human performance. In addition, the benefit of WOC decoding for LLMs/VLMs increases with increasing number of reasoning paths samples. In sum, we introduce guesstimation as a new task that is very common in real world but has been overlooked by the AI community.

Acknowledgements

We thank the reviewers, the area chair for their feedback. This work was funded by the Multi University Research Initiative grant from the Department of Defense, W911NF2110317 (with Rogers as Co-I).

References

[1] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[2] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024.

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

[4] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[5] 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. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[8] Michael Nofer and Michael Nofer. 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, 2015.

[9] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024.

[10] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

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

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

[13] 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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=yzkSU5zdwD`. Survey Certification.

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

[15] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *arXiv preprint arxiv:2404.03622*, 2024.

[16] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

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

[18] Yongqiang Zhao, Zhenyu Li, Zhi Jin, Feng Zhang, Haiyan Zhao, Chengfeng Dou, Zhengwei Tao, Xinhai Xu, and Donghong Liu. Enhancing the spatial awareness capability of multi-modal large language model. *arXiv preprint arXiv:2310.20357*, 2023.

[19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A Guesstimation Questions and Ground Truth Answers

Table 2 lists the guesstimation questions used in the MARBLES dataset along with their corresponding ground truth answers.

Table 2: List of questions and their corresponding true answers.

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

B The Prompts used for querying the LLMs/VLMs

Table 3 lists the prompts that we use when querying the LLMs/VLMs.

Table 3: The prompts used for query the LLMs.

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

C Selection of the LLMs/VLMs

Table 4 lists the LLMs/VLMs that we evaluate.

Table 4: List of large language models.

Condition	Model Family	Model Variant
Verbal	Mistral	mistral-7b-instruct-v0.2
	Mixtral	mixtral-8x7b-instruct-v0.1
		mixtral-8x22b-instruct-v0.1
	LLaMA 2	llama-2-7b-chat-hf
		llama-2-13b-chat-hf
		llama-2-70b-chat-hf
	LLaMA 3.1	llama-3.1-8b
		llama-3.1-70b
	Vicuna	vicuna-7b-v1.5
		vicuna-13b-v1.5
		vicuna-33b-v1.3
	GPT	gpt-3.5-turbo-0125
		gpt-4-0125-preview
Multimodal	LLaVA	llava-v1.6-mistral-7b
		llava-v1.6-vicuna-7b
		llava-v1.6-vicuna-13b