VAQUUM: Are Vague Quantifiers Grounded in Visual Data?

Anonymous ACL submission

Abstract

Vague quantifiers such as *a few* and *many* are influenced by many contextual factors, including how many objects are present in a given context. In this work, we evaluate the extent to which vision-and-language models (VLMs) are compatible with humans when producing or judging the appropriateness of vague quantifiers in visual contexts. We release a novel dataset, VAQUUM, containing 20300 human ratings on quantified statements across a total of 1089 images. Using this dataset, we compare human judgments and VLM predictions using three different evaluation methods. Our findings show that VLMs, like humans, are influenced by object counts in vague quantifier use. However, we find significant inconsistencies across models in different evaluation settings, suggesting that *judging* and *producing* vague quantifiers rely on two different processes.

1 Introduction

Everyday conversations are replete with statements containing vague quantifiers, such as "There are *many* horses" (Figure 1). Despite the fact that they are vague, they cause surprisingly little misunderstanding among interlocutors (Jucker et al., 2003). Vague quantifiers, unlike *crisp* quantifiers, allow for borderline cases in which it is unclear whether the quantifier applies or not, and where we can also expect some variation in the extent to which speakers would use it. For example, *all* does not allow for borderline cases, but it is unclear when a quantity ceases to be *a few* or how many *many* is. Despite the fact that vague quantifiers have long been a subject of investigation among formal semanticists (see e.g. Nouwen, 2010) and (psycho)linguists (e.g. Moxey and Sanford, 1993a; van Deemter, 2010), they have received relatively little attention in the field of natural language processing (NLP).

In visually grounded settings, the use of vague quantifiers can be influenced by factors related to the scene itself, such as the number of entities observed (e.g. Coventry et al., 2005) as well as their sizes (Hörmann, 1983; Coventry et al., 2010), but also by information like the speaker's and hearer's personal beliefs and attitudes (Moxey and Sanford, 2000; Jucker et al., 2003). This broad range of factors, coupled with their vagueness, raises the question of how well computational models of language are able to capture human patterns in the comprehension and use of such expressions. In this paper, we explore this question with vision and language models (VLMs) in multimodal settings involving quantified statements about images. The inclusion of a vision modality allows us to provide context in the form of both visual and textual information (Zhang et al., 2024; Ghosh et al., 2024). Our work follows the spirit of recent research exploring the grounding abilities of VLMs

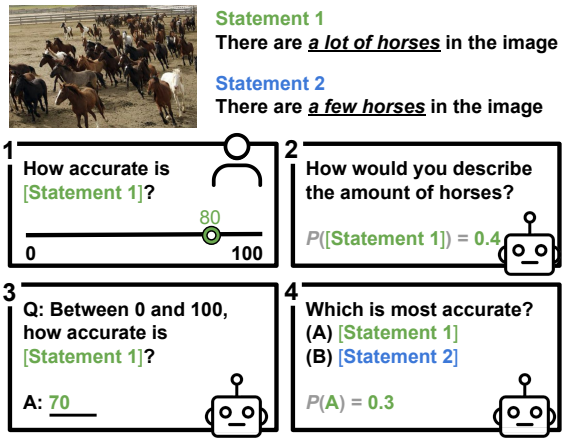

Figure 1: **Experiments in this work.** We (**1**) ask human participants to rate, using a slider, the appropriateness of statements containing vague quantifiers in relation to images. We (**2**) extract VLM generation probabilities for those same statements, (**3**) prompt the models to generate an accuracy score for them and (**4**) evaluate probabilities assigned to these statements in a multiple-choice setup. The image above is originally from the FSC-147 dataset (Ranjan et al., 2021).

(e.g. Zellers et al., 2019; Thrush et al., 2022; Zhang et al., 2022a; Parcalabescu et al., 2022; Chen et al., 2023; Kamath et al., 2024; Wang et al., 2024). We present VAQUUM, a new dataset which pairs images with human judgments on the acceptability of quantified statements. We also examine to what extent visual cues influence state-of-the-art VLMs' understanding and production of expressions containing vague quantifiers, and how this compares to human linguistic intuitions (Figure 1).

The contributions of this paper are as follows.[1]

- We release VAQUUM (**Va**gue **Qu**antifiers with H**um**an Judgments), a new dataset pairing images of different types of objects with their counts, as well as human judgments of different quantified statements corresponding to the image.
- We analyze the features of the visual context that influence both human and model judgments on the appropriateness of different vague quantifiers, including counts, the segmentation area occupied by the target objects, and aspects of world knowledge such as their normative size.
- We show that VLMs do, to some extent, follow human patterns in judging the appropriateness of vague quantifiers, but instruction-tuned models generally align better. However, the behavior of models and their degree of alignment with human judgments depends on the evaluation paradigm used.

2 Related Work

The use and judgment of vague quantifiers has been studied extensively in the psycholinguistics literature. Recent years have also seen a growing but relatively limited interest in studying (V)LM behavior with linguistic quantifiers.

Vague quantifiers in human language Numbers have been shown to play a significant role in the understanding and use of vague quantifiers in humans (see e.g. Solt, 2011). It has been suggested that humans make use of an *approximate number system* (Feigenson et al., 2004; Dehaene, 2011; Coventry et al., 2005), where vague terms might not refer to exact numbers but rather approximations thereof. However, it has also been shown that quantifier comprehension and use go beyond (approximations of) cardinality of the targeted object. Factors include object size (Hörmann, 1983; Newstead and Coventry, 2000); the number and proportions of *other* objects in the scene (Coventry et al., 2005, 2010; Pezzelle et al., 2018); set size (e.g. the answer to a question such as: "*Several* marbles from a set of 12 marbles would be __ marbles"; Newstead et al., 1987); the functionality of objects in the scene (Newstead and Coventry, 2000); and object grouping and spacing (Coventry et al., 2005).

In conversations and texts, the choice of quantifier influences the (expected) rhetorical impact of a statement, and vice versa. Moxey and Sanford (1993b) show that the choice of quantifier can reveal a speaker's *prior expectations* regarding the frequency of the object in the scene. Moreover, several works have outlined the different perspectives that *a few* and *few* convey: while "*a few* people were at the party" focuses on those who were present, "*few* people were at the party" puts the emphasis on those who did not attend (Moxey and Sanford, 2000; Paterson et al., 2009).

(Vague) Quantification with (V)LMs Most work on evaluating VLMs on quantifiers has focused on *crisp* quantifiers (e.g. *none*, *all* and *more than half*) rather than vague ones. Sorodoc et al. (2016) show that neural networks can be trained to learn the quantifiers *no*, *some* and *all* without the need for an explicit counting system. Sorodoc et al. (2018) extend this to a visual question-answering (VQA) task with natural images. They include vague expressions with *few* and *some*, but define these terms using specific proportions (e.g. *few* applies for predications involving less than 17% of objects in the domain). A similar definition is adapted by Pezzelle et al. (2017), who show that models require different mechanisms for learning cardinals and quantifiers. Note that once the range of a quantifier is defined, it can no longer be considered *vague* as borderline cases are excluded.

Moving beyond the gold label approach, Testoni et al. (2019) demonstrate that models using both audio and visual input to select appropriate quantifiers can achieve results that align with human distributions reported by Pezzelle et al. (2018). Enyan et al. (2024) compare human and large language model (LLM) responses on questions such as "There are 500 balls. 234 of them are yellow. Are many balls yellow?" They find that responses generated by LLMs align more closely with human judgments

[1] Code and data will be publicly available.

on crisp quantifiers than on vague ones. Belém et al. (2024) find that LLMs are able to map uncertainty expressions such as *probably* and *unlikely* to probabilistic (numerical) responses in a human-like fashion. More akin to our experiments, Testoni et al. (2024) evaluate three VLMs on their abilities to assign appropriate quantifiers to visual scenes, prompting models to select one out of nine quantifiers in response to questions such as "How many animals are there in the image?", with synthetic images generated by Pezzelle et al. (2018). None of the models show any correlation with the distribution of responses provided by human annotators, which the authors suggest might be due to the models' poor counting abilities. Our approach diverges from theirs on several points. First, we use natural images rather than artificial ones, offering a more realistic setting for evaluating VLMs. Additionally, we use a wider range of methods to provide a more comprehensive assessment of model behavior.

3 The VAQUUM Dataset

We construct the VAQUUM dataset: **Va**gue **Qu**antifiers with **Hum**an Judgments.

Images We utilize annotated datasets used for object counting in computer vision. FSC-147 (Ranjan et al., 2021) contains 6146 images across 147 object types, with annotated object counts ranging from 7 to 3731. Hobley and Prisacariu (2023) refine and deduplicate this dataset to release FSC-133 (containing 133 object types). We sample images from FSC-133 and exclude a total of 22 object categories for several reasons, such as their uncountable nature (e.g. *fresh cut*), obscurity (e.g. *carrom board pieces*) or simply because the images do not depict the object from the label. We also remap 37 categories to either their plural form where necessary or to their basic-level category (e.g. mapping *crows* to *birds*; cf. Rosch et al., 1976). Since the lowest count in FSC-133 is 7, we complement this dataset with samples from the test set of TallyQA (Acharya et al., 2019), which includes images and annotated counts sourced from Visual Genome (Krishna et al., 2017) and VQA2 (Antol et al., 2015; Goyal et al., 2017). Here, we use images classified as "simple" in TallyQA, which have counts between 1 and 15. From this set, we exclude images for which the labelled object is not in the set of remapped FSC-133 labels. We discard all counts below 2 (from TallyQA) and above 100 (from FSC-133). We include three types of object features in our dataset:

1. Count bin To address the imbalance in object counts within the merged dataset, we group the 99 distinct counts (ranging from 2 to 100) into bins of three (counts from 2 to 4, 5 to 7, etc). From each bin, we randomly sample 33 images, yielding 1089 images, evenly distributed across 33 count bins, covering counts from 2 to 100.

2. Segmentation area We estimate the segmentation area of the object(s) in each image, i.e. the ratio of pixels in the objects' bounding region over the total image area. For each image, we prompt CLIPSeg (Lüddecke and Ecker, 2022), with the name of the object type (e.g *birds*). The output logits are than passed through a sigmoid function, and the resulting values are thresholded. The resulting binary mask is used to compute the segmentation area, which essentially corresponds to "object size" in previous work.

3. Size norm We investigate the impact of real-world object size using the object-specific norms in the THINGSplus database (Stoinski et al., 2024), an extension of THINGS (Hebart et al., 2019). Such norms are collected from human judges, and they reflect "average" or "typical" values for specific properties. The *size* norm tells us something about an object's perceived real-life size, on an arbitrary scale. Objects that are not explicitly present in this dataset are either mapped to the closest (base) category or discarded in our size norm analyses.

3.1 Human Judgments

We recruited 203 participants, all native and primary speakers of English, through Prolific (52.2% female; 45.8% male; 1.5% undisclosed). Participant ages ranged from 25 to 84, with the majority aged 25-34 (31.5%) and 35-44 (25.6%).

3.1.1 Procedure

We presented each participant with 100 questions in a random order. Each of these questions consist of an image and a statement of the form "There are `[QUANT]` `[OBJECT]` in the image." Here, `OBJECT` is the plural form of the object depicted and `QUANT` $\in$ {*few*, *a few*, *some*, *many*, *a lot of*} (e.g. "There are *a lot of apples* in the image."). For each image, we also include the unquantified statement (omitting `QUANT`). Participants were asked to rate, using a slider, how accurate the statement is for the image (see Figure 1). The slider ranges from "Completely inaccurate" to "Completely accurate". No participant saw the same image twice.

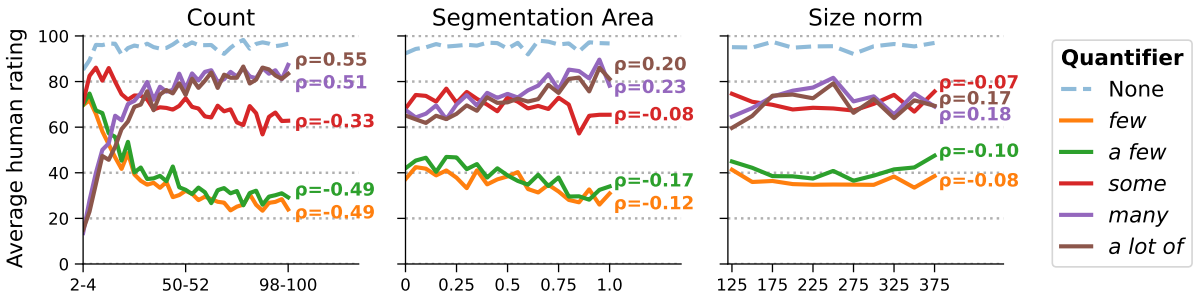

Figure 2: **Average human ratings with increasing counts, segmentation area and size norms.** For each variable and each quantifier, we also report Spearman's ρ, which are all statistically significant ($p < 0.05$).

	few	*a few*	*some*	*many*	*a lot of*	**ME**
C	-0.37	-0.38	-0.20	0.38	0.42	0.03
SG	-0.07	-0.10	-0.05	0.08	0.06	0.04
SN	-0.13	-0.11	-0.07	0.14	0.17	0.01*
ME	-1.71	-1.60	-0.73	-0.60	-0.69	

Table 1: **Estimates of the linear mixed effects model fit to data in VAQUUM. C**=Count, **SG**=Segmentation, **SN**=Size norm, **ME**=Main effect. All numbers are statistically significant ($p < 0.05$), except the one marked (*). For main effects, the quantifier is releveled to the unquantified case, with intercept estimated at $\beta = 0.89$.

3.1.2 Analysis

We analyze the effects of count, segmentation area and size norms on the collected appropriateness ratings of the vague quantifiers. We summarize the results in Figure 2.

We observe from Figure 2 that an increase in count leads to an increase in the average ratings assigned to statements containing *many* and *a lot of*, whose trajectories are nearly identical. Conversely, for the complementary pair *few* and *a few*, we find that average ratings *decrease* as object count increases. As expected, judgments for unquantified control statements are independent of count, with the exception of a slightly lower rating for the lowest counts. We also observe that *few*/*a few* and *many*/*a lot of* exhibit opposing trends in relation to count, again as expected. These observations are broadly in line with findings by e.g. Coventry et al. (2010). Average ratings for *some* also decrease as count increases, though less steeply than for (*a*) *few*. While signs of Spearman's coefficient are the same across all predictors, the strength of the correlation for segmentation area and size norm is noticeably lower. Furthermore, *few*/*a few* and *many*/*a lot of* do not exhibit opposing trends as a function of area or size norm, as they do with count.

To gain further insights into the relations between participants ratings and object count and size, we fit a linear mixed effects model (LMM) to our data, predicting human judgments from the fixed effects of quantifiers, count, segmentation area and size norm and using participants and object category as random effects. We include interaction terms between pairs of predictors to investigate their joint influence on judgments. For full details of the LMM, we refer to Appendix B.

We report LMM estimates of the main effects and two-way interaction effects in Table 1. All main effects except those for size norm are statistically significant. For the two-way interactions, *few*, *a few* and *some* consistently show negative estimates across all predictors, while *many* and *a lot of* are consistently positive. As expected given the trends in Figure 2, object count exhibits the strongest impact on each quantifier. Estimates for segmentation area and size norm display similar trends, but with weaker effects. The LMM explains 50.3% of the total variance in our participant data ($R^2c = 0.503$, $R^2m = 0.459$). The random effects present moderate variability at participant level, with a variance of 0.042 suggesting that individual differences among participants explain some of the variance in judgments. In contrast, the object random effect accounts for minimal variance (0.002), indicating that differences between objects have little influence on the judgments given by participants in our experiments.

4 Experiment 1: Production Probabilities

Our first series of experiments studies the predicted *production probabilities* of quantified statements by SOTA VLMs. We prompt the models with "How would you describe the amount of `[OBJECT]` in the image?" We extract log probabilities, conditioned on this prompt and the image, for the quantified

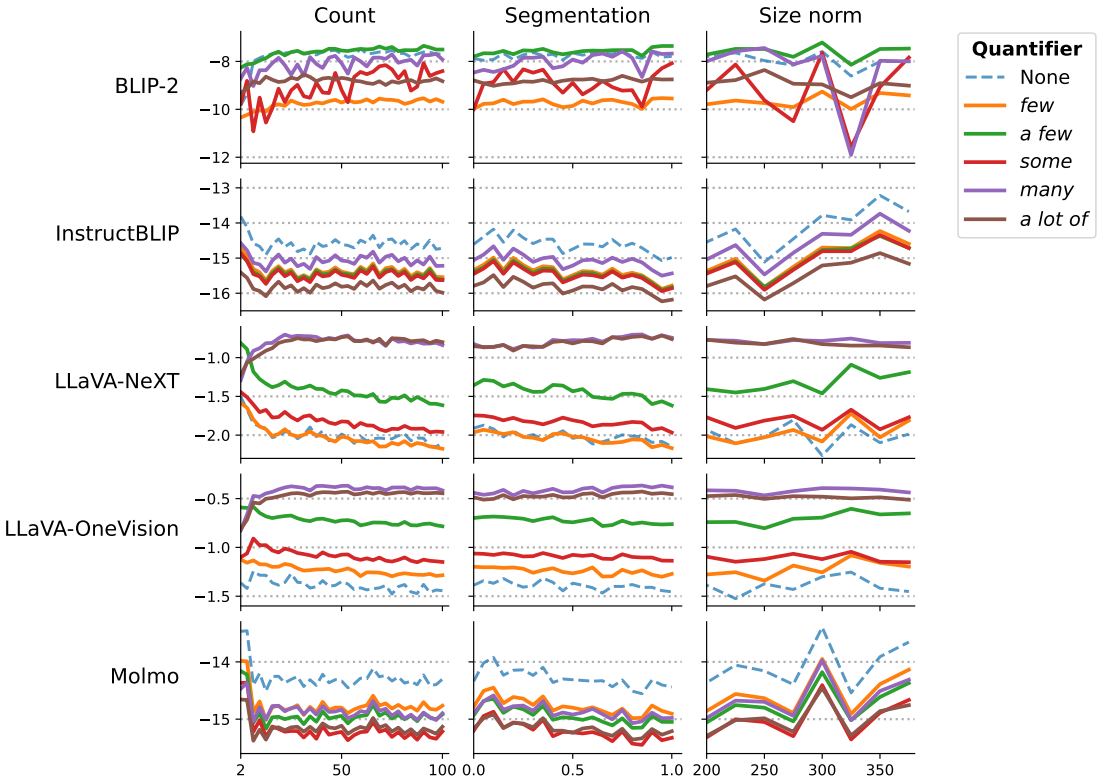

Figure 3: **Log probabilities as functions of count, segmentation area and size norm.** The patterns reported for LLaVA-NeXT and LLaVA-OneVision are most similar to human ratings. We find that InstructBLIP and Molmo do not distinguish between the quantifiers at all, while BLIP-2 moderately correlates with humans for *many/a lot of*.

statements in VAQUUM, as well as the unquantified version. All extracted scores are normalized by token length. We consider the following models.

BLIP-2 (Li et al., 2023)**.** We use the checkpoint powered by OPT-6.7B (Zhang et al., 2022b) connected to a EVA-CLIP ViT-g (Radford et al., 2021; Fang et al., 2023) image encoder via a lightweight Query transformer.

InstructBLIP (Dai et al., 2023)**.** We use the checkpoint with a Vicuna-13B (Zheng et al., 2023) language backbone, instruction-tuned on BLIP-2.

LLaVA-NeXT (Liu et al., 2024)**.** We use the 7B checkpoint with a Mistral (Jiang et al., 2023) language backbone.

LLaVA-OneVision (Li et al., 2024)**.** We utilize the 7B checkpoint, which integrates a SigLIP (Zhai et al., 2023) vision encoder with a Qwen2 (Yang et al., 2024) language decoder.

Molmo (Deitke et al., 2024)**.** We use the 7B-D checkpoint, which connects a ViT image encoder to Qwen2.7B via a connector MLP.

Figure 3 displays predicted log probabilities as a function of count, segmentation area and size norm and Table 2 reports correlations between model predictions and human judgments.

Alignment with humans Of the VLMs tested, the two LLaVA models exhibit the highest correlation with the human data in VAQUUM. For these models, we observe in Figure 3 patterns similar to those of VAQUUM in Figure 2. Probabilities for *many* and *a lot of* increase as a function of count, while *few* and *a few* show a downward trend. Given that the question in the prompt focused explicitly on the *amount* of objects, the unquantified sentence is expected to be generally dispreferred. The trends in Figure 3 suggest that the LLaVA models can indeed draw this distinction between quantified and unquantified statements, as the unquantified expression displays lowest-ranking log probabilities across count, segmentation and size norm. However, other models do not reveal that same ability. This is most pronounced for InstructBLIP and Molmo, which generally tend to favor the unquantified statement as a response to the question. These models also show the same pattern across all quantifiers, further confirming their inability to differentiate among them. While the behavior of BLIP-2 is seemingly random, Figure 3 shows an upward trend for all quantifiers as a function of count.

Model	*few*	*a few*	*some*	*many*	*a lot of*
BLIP-2	**-0.18**	**-0.19**	**-0.06**	**0.14**	**0.13**
InstBLIP	0.06	0.04	-0.03	-0.01	-0.04
LLaVA-N	**0.34**	**0.39**	**0.21**	**0.43**	**0.52**
LLaVA-O	**0.30**	**0.40**	**0.22**	**0.52**	**0.54**
Molmo	**0.16**	**0.20**	**0.07**	**-0.17**	**-0.21**

Table 2: **Pearson's correlation between human ratings and model log probabilities.** Numbers in boldface are statistically significant ($p < 0.05$).

	few	*a few*	*some*	*many*	*a lot of*	ME
C	0.00	-0.01	-0.02	**0.22**	**0.22**	**-0.09**
SG	-0.02	-0.01	0.01	**0.07**	**0.05**	**-0.05**
SN	**0.04**	**0.05**	-0.03	**0.12**	**0.09**	**-0.05**
ME	**0.39**	**1.68**	**0.77**	**2.46**	**2.32**	

Table 3: **Estimates of the LMM for log probabilities of LLaVA-OneVision**. **C**=Count, **SG**=Segmentation, **SN**=Size norm, **ME**=Main effect. Boldface indicates statistical significance ($p < 0.05$). For the main effects, the quantifier is releveled to the unquantified case and the estimate of the intercept is $\beta =$ **-1.25**.

Linear mixed model In Table 3, we display the estimates of a linear mixed effects model fit to log probabilities of LLaVA-OneVision (see Appendix B for details and Appendix C for the remaining models). Following our approach in §3, we predict model probabilities from the fixed effects of quantifiers, count, segmentation area and size norm while including object category as a random effect. The latter shows a variance of 0.056, indicating that object category accounts for a moderate amount of variance among predicted log probability scores. Moreover, we see in Table 3 that *many* and *a lot of* show statistically significant interactions with all predictors, with the strongest effects observed with count, just as was the case for the human judgments. The estimates for the other quantifiers, however, are very different from what we found for humans. Overall, the LMM explains 91.2% of the total variance in our data ($R^2m = 0.861$, $R^2c = 0.912$).

Prompts should target *amounts* For most models, we find that simply changing the question from "How would you describe *the amount of* `[OBJECT]` in the image?" to "How would you describe the image?" yields different patterns in the results (see Appendix C). Most notably, we find that the observed similarity between trends in human judgments and model predictions disappear once the prompt does not focus on amounts.

Interim conclusion In §3.1, most estimates of the LMM fit to participant data were statistically significant. Moreover, object count made the biggest difference across all quantifiers. For LLaVA-OneVision, the model displaying the highest Pearson's correlation with human data in Table 2, a similar result can be found in Table 3 for *many* and *a lot of*: effects of interaction with object count are most pronounced, after which size norms have a slightly higher impact than segmentation area. However, these effects are absent for the other quantifiers. BLIP-2, InstructBLIP and Molmo do not show meaningful interactions between their predicted log probabilities and the three contextual variables.

5 Experiment 2: Generating Judgments

We now evaluate the instruction-tuned VLMs using an approach that is more akin to the way VAQUUM was constructed in §3. That is, we prompt the models to explicitly rate the acceptability of quantified statements. We experimented with 10 different prompts that are variations on the question shown to human participants in §3.1. Drawing inspiration from prompts used by Belém et al. (2024), we center our analyses in the remainder of this section around the following prompt: "On a scale of 0 (completely inaccurate) to 100 (completely accurate), how accurate is the following statement for the image? Please respond with one of the following options: 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100. `[Statement]`", where `Statement` is an expression from the VAQUUM dataset. We refer to Appendix D for the complete list of prompts tested.

For VLMs, appropriateness is not gradable We find that in this evaluation setup, BLIP-2 and InstructBLIP generally fail to generate numerical responses to the prompts we tested, despite some prompts explicitly encouraging them to only respond with a number. The two LLaVA models and Molmo consistently provide numerical responses to most of the prompt templates tested. However, while we construct the prompts in such a way that VLMs are encouraged to provide a response that falls *between* a certain range, the vast majority of model responses tend towards the extremes (i.e. on the lower or upper bound of the specified range; see Appendix D for a distribution of responses).

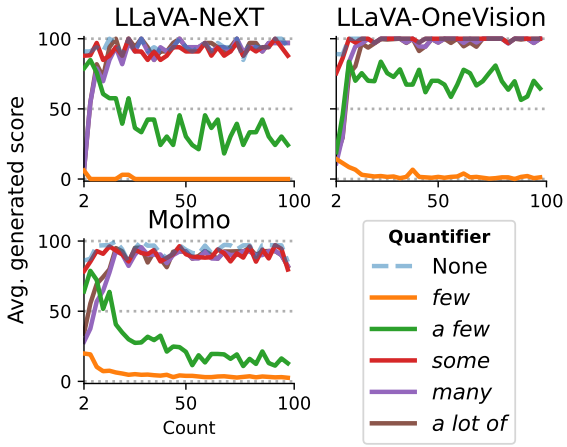

Figure 4: **Scores generated by VLMs in Experiment 2.** Note that we do not display results for BLIP-2 and InstructBLIP, as those models generally failed to provide numerical responses to the prompt.

***Some* is generally appropriate** When numeric answers to prompts tend towards the extremes of a scale, it can be informative to aggregate generated scores, which is virtually the same as calculating the *relative frequency* of a VLM dis/agreeing with the statements. We report this in Figure 4 for object count and make the following observations. First, statements containing the quantifier *few* are rarely deemed appropriate. For the models in the LLaVA family, arguably the most interesting deviation from Figure 3 is that in this setting, *some* is considered an accurate quantifier, regardless of object count. Indeed, we observe that the trajectory of *some* in Figure 4 corresponds to that of the unquantified condition, for which the statements are, as anticipated, generally accepted. We hypothesize that in the case of *judging* the appropriateness of *some*, this vague quantifier could be interpreted as an *existential* quantifier. That is, "There are some apples in the image" can be regarded as a confirmation of the existence of apples in the image.

Interim conclusion Experiment 1 showed that object count has an influence on model predictions for *many* and *a lot of*. Similar patterns emerge in Figure 4, where average scores for these quantifiers increase with count. Discrepancies between results from Experiments 1 and 2 show that in a setting where models are explicitly required to provide judgments for statements (Exp 2), the outcomes are unrelated to the models' log probabilities for the same statements (Exp 1). In Experiment 1, log probabilities are extracted using an autoregressive method compatible with the pretraining objective of the LLM backbone of a VLM. In contrast, Experiment 2 relies on model abilities acquired during post-training, which further modifies model parameters. The discrepancies we observe align with independent observations that post-training can negatively impact model calibration (Kalai and Vempala, 2024; Zhu et al., 2023).

6 Experiment 3: Multiple-Choice QA

Finally, we evaluate VLM judgments in a multiple-choice question-answering (MCQA) setup using a standard MCQA template of the form "Question: Which statement is most accurate for the image? Select the answer from the options below. `[OPTIONS]` Answer: (", with `OPTIONS` being the set of all statements for an image in VAQUUM, labeled `(A)` to `(F)`. For each image, the order of the expressions is shuffled to mitigate the effects of positional biases (Zong et al., 2024). To compare the different quantifiers and ensure that the VLMs do not produce irrelevant output, we extract the log probabilities of the labels rather than allowing VLMs to generate a response. Note that, differently from §4 and §5, the VLMs are now presented with *all* statements before being prompted for a response.

In Figure 5, we report the predicted log probabilities of *instruction-tuned* VLMs as a function of count. Table 4 shows the correlation of these scores with both the human judgments and the log probabilities from Experiment 1. It is clear that in this setup, too, InstructBLIP fails to differentiate between the various quantified statements. However, while Molmo behaved similarly in Experiment 1, it distinguishes between quantifiers in the current setting. For Molmo and the two LLaVA models, count influences predictions for *many/a lot of* and for *few/a few* in the expected direction. This is most pronounced in the lower count ranges. Patterns for *some* once again differ from those found in our earlier experiments. While probabilities for *some* generally fell between those of *few* and *a few* in Experiment 1, and *some* was generally judged appropriate in Experiment 2, we now observe that it follows the same trend as *few* and *a few*, while being slightly preferred over these two by LLaVA-OneVision.

Interim conclusion The two LLaVA models and Molmo show moderate correlation with human scores in VAQUUM. They also correlate with their

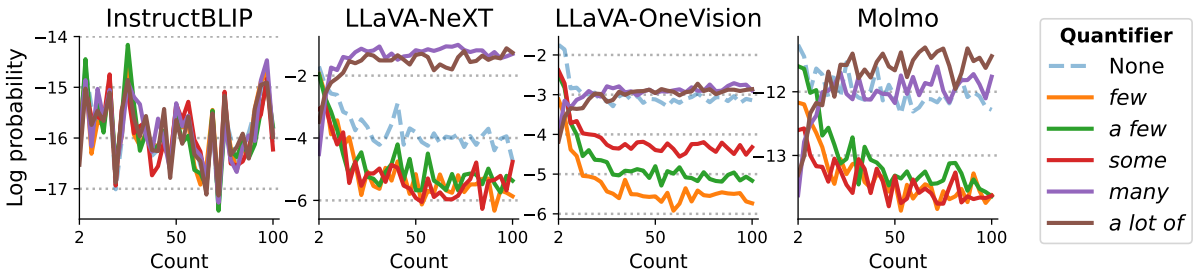

Figure 5: **Log probabilities extracted for multiple-choice labels in Experiment 3.** We do not display results for BLIP-2 because that model is not instruction-tuned.

		few	*a few*	*some*	*many*	*a lot of*
INB	r(VAQ)	0.00	0.00	0.01	-0.01	0.04
	r(EXP1)	**-0.13**	**-0.14**	**-0.12**	**-0.13**	**-0.15**
LLN	r(VAQ)	**0.32**	**0.27**	**0.14**	**0.42**	**0.33**
	r(EXP1)	**0.36**	**0.35**	**0.26**	**0.44**	**0.35**
LLO	r(VAQ)	**0.45**	**0.45**	**0.19**	**0.35**	**0.43**
	r(EXP1)	**0.33**	**0.42**	**0.24**	**0.35**	**0.42**
MOL	r(VAQ)	**0.26**	**0.31**	**0.15**	**0.28**	**0.35**
	r(EXP1)	**0.25**	**0.28**	**0.25**	**-0.07**	**-0.12**

Table 4: **Pearson's r of log probabilities in Experiment 3 with human data (VAQ) and log probabilities from Experiment 1 (EXP1).** Models shown are InstructBLIP (**INB**), LLaVA-NeXT (**LLN**), LLaVA-OneVision (**LLO**) and Molmo (**MOL**). Boldfaced numbers are statistically significant.

log probabilities from Experiment 1. These models are also the most self-consistent. While Molmo is not self-consistent, in the multiple-choice setup it correlates better with human ratings.

7 Discussion

Alignment with humans In this paper, we explore how vision-and-language models produce and evaluate simple expressions containing vague quantifiers. We constructed the VAQUUM dataset and used this to investigate whether object count, segmentation area and size norm affect VLMs to the same extent as they do humans. We showed that in particular for object count, the patterns found in some VLMs show striking similarities with the human data in VAQUUM. This result appears to contradict the observation that VLMs perform poorly on counting tasks (Parcalabescu et al., 2021, 2022). However, our findings with vague quantifiers could be accounted for in terms of an *approximate number system*, which cognitive scientists have posited to account for the human ability to rapidly estimate quantities (Feigenson et al., 2004; Condry and Spelke, 2008; Dehaene, 2011; Odic and Starr, 2018; Piantadosi, 2016). In the context of vague quantifiers, it has been argued that there exists a mapping between exact and approximate number systems (Coventry et al., 2005, 2010). The extent to which VLMs rely on something akin to an ANS is a topic for future work.

Self-consistency Our experiments relied on paradigms incorporating *production* (Experiment 1) and *judgment* (Experiments 2 and 3). We find that VLMs are *not* self-consistent across these evaluation paradigms. That is to say, when a VLM is set to *judge* the use of a quantifier—a meta-linguistic task—its judgment is not rooted in the log probabilities that govern the model's generation of the quantifier.

Outlook Psycholinguistics has shown that vague quantifiers do not depend exclusively on the count and size of target objects. This is further confirmed by the residual variance (49.7%) in VAQUUM that cannot be explained by the linear mixed effects model (LMM) on human judgments. While the LMM analysis yields a better fit for VLM log probabilities, we find that there, too, the LMM cannot explain all the variance (leaving a residual variance of 8.8% for LLaVA-OneVision). Future work could focus on other contextual factors, such as the number of *other* objects present, the object density in the image, as well as the role of scene semantics and other objects in the image background. In combination with visual grounding capabilities, it is worthwhile to investigate the role of commonsense and world knowledge in vague quantifier usage: while seeing 20 people at a conference will most likely not be reason for one to exclaim that there are *many*, the same amount of toddlers at such an event might be.

Limitations

Model selection Our experiments focus on a selection of vision-and-language models. While this selection has allowed us to compare models from the same model family (BLIP-2 and InstructBLIP; LLaVA-NeXT and LLaVA-OneVision), as well as models that share similar language model backbones (LLaVA-OneVision and Molmo), conclusions drawn in this study can be better generalised with experiments on a wider range of VLMs. We hope that VAQUUM provides the impetus for further model comparisons.

Segmentation area and size norm Given that of the three contextual variables, the role of object count has been most prominent in literature on vague quantifiers, we focused on selecting images that balance a range of counts that we deemed representative. Estimating the segmentation area and extracting the size norms for these images may subsequently have yielded distributions that do not represent the full range of values that these variables can take on. It is therefore possible that the distributions for segmentation area and size norm were too sparse to say something more meaningful about their roles in VAQUUM and model results. Thus, while we at times find statistically significant relationships between judgments and segmentation area or size norm, future work could focus on investigating the *practical* significance. Additionally, we recognize that using CLIPSeg to estimate the segmentation area can introduce inaccuracies.

Variance in human judgments By aggregating human judgments through simply taking the average and focusing on general trends, we might overlook meaningful variability that emphasize the complexity of human judgments on vague expressions. While the aim of this work was to investigate whether VLMs can approximate general patterns in human data, we believe that VAQUUM is a dataset that can contribute to the study of disagreement among human annotators.

Ethical Considerations

The data collection study for VAQUUM underwent an ethics check in our institution. The data collected via crowdsourcing does not contain any information that can be traced back to individuals. No materials were used to our knowledge which could harm or otherwise adversely affect individuals.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Catarina G Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8467–8502, Miami, Florida, USA. Association for Computational Linguistics.

Xinyi Chen, Raquel Fernández, and Sandro Pezzelle. 2023. The BLA Benchmark: Investigating Basic Language Abilities of Pre-Trained Multimodal Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5817–5830, Singapore. Association for Computational Linguistics.

Kirsten F Condry and Elizabeth S Spelke. 2008. The development of language and abstract concepts: the case of natural number. *Journal of Experimental Psychology: General*, 137(1):22.

Kenny R. Coventry, Angelo Cangelosi, Stephen E. Newstead, and Davi Bugmann. 2010. Talking about quantities in space: Vague quantifiers, context and similarity. *Language and Cognition*, 2(2):221–241.

Kenny R Coventry, Stephen Newstead, and Rohanna Rajapakse. 2005. Grounding Natural Language Quantifiers in Visual Attention. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Stanislas Dehaene. 2011. The number sense: How the mind creates mathematics.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh,

Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Zhang Enyan, Zewei Wang, Michael A. Lepori, Ellie Pavlick, and Helena Aparicio. 2024. Are LLMs models of distributional semantics? a case study on quantifiers. *Preprint*, arXiv:2410.13984.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, Los Alamitos, CA, USA. IEEE Computer Society.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *Preprint*, arXiv:2404.07214.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24.

Michael Hobley and Victor Prisacariu. 2023. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hans Hörmann. 1983. *The Calculating Listener, or: How Many are einige, mehrere and ein paar (Some, Several, and a Few)?*, pages 221–234. De Gruyter, Berlin, Boston.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Andreas H. Jucker, Sara W. Smith, and Tanja Lüdge. 2003. Interactive aspects of vagueness in conversation. *Journal of Pragmatics*, 35(12):1737–1769.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*. Association for Computing Machinery.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The Hard Positive Truth about Vision-Language Compositionality. *arXiv preprint*. ArXiv:2409.17958 [cs].

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096.

Linda M Moxey and Anthony J Sanford. 1993a. *Communicating quantities: A psychological perspective.* Lawrence Erlbaum Associates, Inc.

Linda M. Moxey and Anthony J. Sanford. 1993b. Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology*, 5(1):73–91.

Linda M. Moxey and Anthony J. Sanford. 2000. Communicating quantities: a review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*, 14(3):237–255. Publisher: Wiley-Blackwell.

S.E. Newstead, P. Pollard, and D. Riezebos. 1987. The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, 18(3):178–182.

Stephen E. Newstead and Kenny R. Coventry. 2000. The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology*, 12(2):243–259.

Rick Nouwen. 2010. What's in a quantifier? In Martin B.H. Everaert, Tom Lentz, Hannah N.M. De Mulder, Øystein Nilsen, and Arjen Zondervan, editors, *The Linguistic Enterprise*, Linguistik Aktuell 150, pages 235–256. John Benjamins.

Darko Odic and Ariel Starr. 2018. An introduction to the approximate number system. *Child Development Perspectives*, 12(4):223–229.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 32–44, Groningen, Netherlands (Online). Association for Computational Linguistics.

Kevin B. Paterson, Ruth Filik, and Linda M. Moxey. 2009. Quantifiers and discourse processing. *Language and Linguistics Compass*, 3(6):1390–1402.

Sandro Pezzelle, Raffaella Bernardi, and Manuela Piazza. 2018. Probing the mental representation of quantifiers. *Cognition*, 181:117–126.

Sandro Pezzelle, Marco Marelli, and Raffaella Bernardi. 2017. Be precise or fuzzy: Learning the meaning of cardinals and quantifiers from vision. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 337–342, Valencia, Spain. Association for Computational Linguistics.

Steven T. Piantadosi. 2016. A rational analysis of the approximate number system. *Psychonomic Bulletin & Review*, 23(3):877–886.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. 2021. Learning To Count Everything . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3393–3402, Los Alamitos, CA, USA. IEEE Computer Society.

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in Natural Categories. *Cognitive Psychology*, 439:382–439.

Stephanie Solt. 2011. Vagueness in Quantity: Two Case Studies from a Linguistic Perspective. In *Understanding vagueness. Logical, philosophical and linguistic perspectives*. College Publications.

Ionut Sorodoc, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle, and Raffaella Bernardi. 2016. "Look, some green circles!": Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 75–79, Berlin, Germany. Association for Computational Linguistics.

Ionut Sorodoc, Sandro Pezzelle, Aurélie Herbelot, Mariella Dimiccoli, and Raffaella Bernardi. 2018. Learning quantification from images: A structured neural architecture. *Natural Language Engineering*, 24(3):363–392.

Laura M Stoinski, Jonas Perkuhn, and Martin N Hebart. 2024. THINGSplus: New norms and metadata for the things database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, 56(3):1583–1603.

Alberto Testoni, Sandro Pezzelle, and Raffaella Bernardi. 2019. Quantifiers in a multimodal world: Hallucinating vision with language and sound. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 105–116, Minneapolis, Minnesota. Association for Computational Linguistics.

Alberto Testoni, Juell Sprott, and Sandro Pezzelle. 2024. Naming, describing, and quantifying visual objects in humans and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 547–557, Bangkok, Thailand. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5228–5238.

Kees van Deemter. 2010. *Not Exactly: in Praise of Vagueness*. Oxford University Press.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024. Mementos: A Comprehensive Benchmark for Multimodal Large Language Model Reasoning over Image Sequences. *arXiv preprint*. ArXiv:2401.10529 [cs].

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan

Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5321–5335, Seattle, United States. Association for Computational Linguistics.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. OPT: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Preprint*, arXiv:2306.05685.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the Calibration of Large Language Models and Alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore. Association for Computational Linguistics.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. 2024. Fool Your (Vision and) Language Model With Embarrassingly Simple Permutations. *arXiv preprint*. ArXiv:2310.01651.

A Data from Human Participants

A.1 Instructions and Consent

Below we include the information given to the participants in our human experiment.

Thank you for taking part in this experiment. This survey should take approximately 20 minutes to complete. You will be presented with 100 questions. Each question consists of an image and a corresponding statement. Your task is to rate, using a slider, how accurate you find the statement in relation to the image.

Please be assured that all responses will be kept strictly confidential and anonymous. The data that we collect will be processed in such a way that they cannot be linked to you in any way. Participation in this survey is entirely voluntary. If at any point you wish to exit the survey without finishing the survey, you can close this form and we will delete your responses. You do not have to specify your reason.

Should you wish to withdraw consent after you have participated, please send an email to **AUTHORS** *at* **EMAIL**. *Note that if you withdraw consent after completing the survey, we are not required to undo the processing of your data that has taken place up until that time.*

If you wish to participate in the study, please check the following box. If you do not wish to do so, you can close this tab.

A.2 Demographics

In §3.1, we mentioned that we recruited 203 participants through Prolific. As reported in the Ethical Considerations, we did not collect data that allows anyone to trace the responses back to an individual. All participants were native and primary speakers of English. Besides that fact, we have the following information about the distribution of demographic information.

Age 25-34 years (31.5%), 35-44 (25.6%), 18-24 (17.2%), 45-54 (15.3%), 55-64 (6.9%), 65-74 (2.5%) and 75-84 (0.5%). 0.5% of the participants prefer not to disclose their age.

Gender female (52.5%), male (45.8%), other (0.5%). 1.5% of the participants prefer not to say.

A.3 Participant reward

Participants were found through Prolific and were paid £ 2.50 for 20 minutes (£ 7.50 per hour).

B Linear Mixed Effects Models

Below we provide the details for the linear mixed effects models that we fit to our data. All LMMs are fit using the `lme4` package in R.

B.1 Human Data (VAQUUM)

In §3, we are interested in predicting human judgments from the main effects of quantifiers, object count, segmentation area and size norms, as well as the interaction between these predictors. We include the participants and object categories as random effects. Put concretely,

```
judgment ~ quantifier * count
         * segmentation * size_norm
         + (1|participant) + (1|object)
```

We scale judgments, count, segmentation area and size norm to make sure they all have a mean of 0 and a standard deviation of 1. For example,

```
count <- scale(count,
               center=TRUE,
               scale=TRUE)
```

This way, we ensure that we can meaningfully interpret the relation between one unit of change in one variable with a change in another. Additionally, we make the variables for quantifier and object category a `factor` and relevel the quantifier to use the unquantified (`base`) condition as the reference category.

```
quantifier <- relevel(quantifier,
                      ref="base")
```

B.2 Model Data (Experiment 1)

For the models, we follow the same steps taken as those for fitting an LMM to human data, but now we no longer have to account for different participants. That is,

```
log_prob ~ quantifier * count
         * segmentation * size_norm
         + (1|object)
```

C Supplementary Material Experiment 1

C.1 Targeting amounts

In Figure 6 we show the patterns of the VLMs across all predictors for the prompt that does *not* target the amount. The question presented to the models is "How would you describe the image?", and we extract log probabilities for expressions of the form "There are `[QUANT]` `[OBJECT]` in the image" (unchanged from those used in §4).

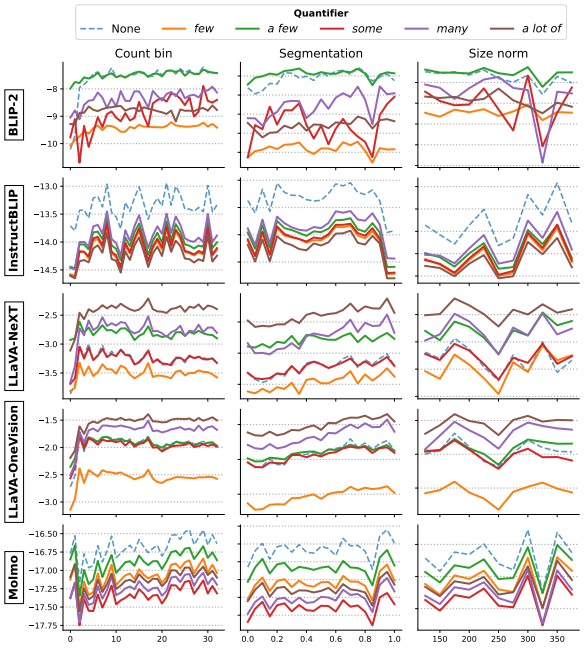

Figure 6: **Log probabilities extracted for statements as a response to "How would you describe the image?"** The most obvious deviation from Figure 3 in §4 are the plots for the two LLaVA models, that no longer appear to distinguish between the different quantified statements.

For LLaVA-NeXT and LLaVA-OneVision, the two models observed in §4 to have the highest correlation with human ratings, we now find that patterns are the same across all quantifiers. We now find a "layered" or "stacked" pattern that is indicative of a bias towards a specific quantifier: while LLaVA-NeXT and LLaVA-OneVision tend towards always responding with *a lot of*, InstructBLIP and Molmo favor the unquantified statement.

C.2 LMMs for all remaining models

In Table 5, we report estimates of LMMs for BLIP-2, InstructBLIP, LLaVA-NeXT and Molmo.

D Supplementary Material Experiment 2

D.1 Prompts for Score Generation

Below we list the 10 prompts that we have tested for Experiment 2. The prompt listed in boldface is discussed in §5.

1. "On a scale of 0 (completely inaccurate) to

		Intercept	Main	Quantifier				
				few	*a few*	*some*	*many*	*a lot of*
BLIP-2	**Main effect**	**0.41**	–	**-0.89**	**-0.09**	**-0.79**	**-0.26**	**-1.37**
	Count		0.03	**0.21**	0.02	**-0.10**	-0.03	0.01
	Segmentation		0.03	**0.06**	-0.04	-0.02	**0.09**	-0.02
	Size norm		0.02	0.01	-0.03	**-0.07**	**-0.13**	0.06
InstructBLIP	**Main effect**	**0.57**	–	**-0.76**	**-0.82**	**-0.86**	**-0.46**	**-1.20**
	Count		-0.02	-0.02	-0.01	-0.01	0.03	0.00
	Segmentation		**-0.11**	-0.01	0.00	0.00	0.00	0.02
	Size norm		**0.33**	**-0.08**	**-0.09**	**-0.06**	0.02	**-0.09**
LLaVA-NeXT	**Main effect**	**-0.86**	–	**-0.05**	**1.00**	**0.31**	**2.10**	**2.08**
	Count		**-0.12**	**-0.03**	**-0.07**	**-0.04**	**0.21**	**0.26**
	Segmentation		**-0.12**	0.00	-0.03	0.02	**0.14**	**0.13**
	Size norm		**-0.08**	**0.08**	**0.12**	**0.03**	**0.15**	**0.15**
LLaVA-OneVision	**Main effect**	**-1.25**	–	**0.39**	**1.68**	**0.77**	**2.46**	**2.32**
	Count		**-0.09**	0.00	-0.01	-0.02	**0.22**	**0.22**
	Segmentation		**-0.05**	-0.02	-0.01	0.01	**0.07**	**0.05**
	Size norm		**-0.05**	**0.04**	**0.05**	-0.03	**0.12**	**-0.09**
Molmo	**Main effect**	**0.73**	–	**-0.71**	**-0.97**	**-1.35**	**-0.85**	**-1.30**
	Count		**-0.11**	0.03	0.03	**-0.05**	0.02	0.04
	Segmentation		**-0.19**	0.00	0.00	-0.01	0.02	0.03
	Size norm		**0.22**	-0.01	-0.04	-0.05	0.01	**-0.06**

Table 5: **Linear Mixed Effects estimates for all VLMs tested.** We discuss the estimates for LLaVA-OneVision in §4.

100 (completely accurate), how accurate is the following statement for the image? Please only respond with a number between 0 and 100.
`[Statement]`"

2. "Question: How accurate is the following statement for the image? Respond only with a rating between 0 (completely inaccurate) and 100 (completely accurate).
Statement: `[Statement]`
Answer: "

3. "On a scale of 0 (completely inaccurate) to 100 (completely accurate), how accurate is the following statement for the image? Respond only with a number. Decimals are allowed.
`[Statement]`"

4. "How accurate is the statement for the image? Please only respond with a number between 0 and 100, where 0 is 'completely inaccurate' and 100 'completely accurate'.
`[Statement]`"

5. "**On a scale of 0 (completely inaccurate) to 100 (completely accurate), how accurate is the following statement for the image? Please respond with one of the following options: 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100.**
`[Statement]`"

6. "How likely is the following caption given the image? Please respond with a number between 0 and 100, where
- 0 is 'not likely at all'
- 100 is 'highly likely'.
Caption: `[Statement]`"

7. "What is the probability that the following sentence matches the image?
`[Statement]`"

8. "What is the probability that the following sentence matches the image?
Sentence: `[Statement]`
Answer: "

9. "What is the probability that the following sentence matches the image? Please only respond with a number between 0 and 100.
`[Statement]`"

10. "What is the probability that the following sentence matches the image? Please only respond with a number between 0 and 1.
Sentence: `[Statement]`
Answer: "

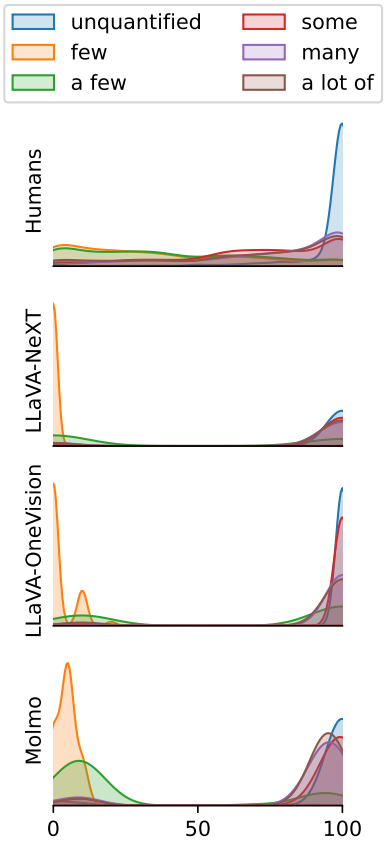

Figure 7: **Distributions for human ratings and scores generated by VLMs per quantifier.**

D.2 Distribution of Generated Scores

Figure 7 shows density plots displaying the distributions of human ratings in VAQUUM, as well as scores generated by VLMs as a response to prompt 5 in Appendix D.1, discussed in §5. Note that for LLaVA-NeXT, LLaVA-OneVision and Molmo, the scores tend towards the extremes. However, in the human distribution, this is only the case for the unquantified control statement (as expected).

E Dataset Licenses

For the construction of the VAQUUM dataset, we have used images from existing datasets. We list their licenses below.

TallyQA Apache License 2.0.

FSC-147 MIT License.

FSC-133 is MIT License.

Visual Genome Creative Commons Attribution 4.0 International License

VQA and VQA2 Commons Attribution 4.0 International License

The way we include these datasets in our experiments is consistent with their intended use.