

Improving Automatic VQA Evaluation Using Large Language Models

Oscar Mañas^{1,2}, Benno Krojer^{1,3}, Aishwarya Agrawal^{1,2}

¹Mila

²Université de Montréal

³McGill University

oscar.manas@mila.quebec

Abstract

8 years after the visual question answering (VQA) task was proposed, accuracy remains the primary metric for automatic evaluation. VQA Accuracy has been effective so far in the IID evaluation setting. However, our community is undergoing a shift towards open-ended generative models and OOD evaluation. In this new paradigm, the existing VQA Accuracy metric is overly stringent and underestimates the performance of VQA systems. Thus, there is a need to develop more robust automatic VQA metrics that serve as a proxy for human judgment. In this work, we propose to leverage the in-context learning capabilities of instruction-tuned large language models (LLMs) to build a better VQA metric. We formulate VQA evaluation as an answer-rating task where the LLM is instructed to score the accuracy of a candidate answer given a set of reference answers. We demonstrate the proposed metric better correlates with human judgment compared to existing metrics across several VQA models and benchmarks. We hope wide adoption of our metric will contribute to better estimating the research progress on the VQA task. We plan to release the evaluation code and collected human judgments.

Introduction

Visual question answering (VQA) (Antol et al. 2015) has become an essential benchmark for assessing the progress of multimodal vision-language systems. 8 years after the task was proposed, accuracy remains the primary metric for automatically evaluating model performance. VQA Accuracy is based on exact string matching between a candidate answer predicted by the model and a set of reference answers annotated by humans. As pointed out in Agrawal et al. (2023), this metric has been effective so far because the VQA evaluation primarily followed the the independent and identically distributed (IID) paradigm, where the training and testing data distributions are quite similar. Thus, models could learn to adapt to the test answer distribution. However, recently, our community has been shifting its focus towards out-of-distribution (OOD) evaluation, either via zero-shot transfer to unseen VQA tasks or via finetuning on one VQA dataset and evaluating on another (Agrawal et al. 2023). In these settings, the answers generated by VQA models might not

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
arXiv version: <https://arxiv.org/abs/2310.02567>

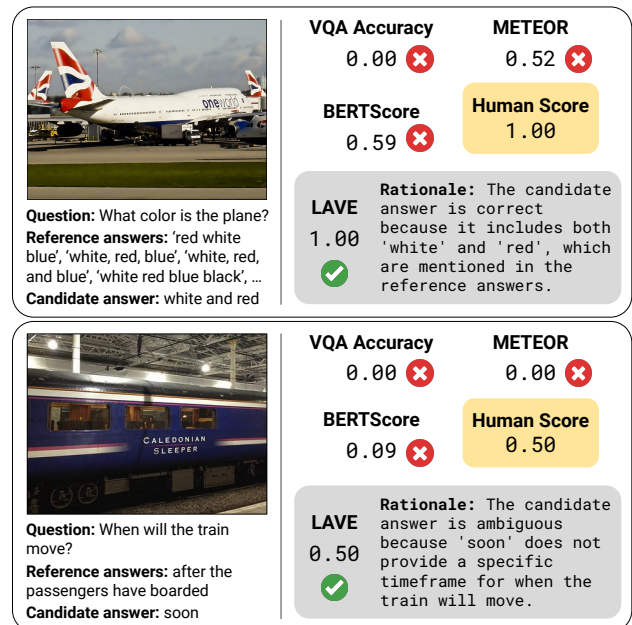


Figure 1: Existing VQA metrics and other strong baselines tend to miss out on correct answers generated by VQA models. Our proposed metric, LAVE, is more aligned with human judgment and provides a rationale for its rating, making it also more interpretable.

match any of the reference answers, while still being correct answers to the question! For instance, the generated answer might differ from the reference answers due to the format, specificity, different interpretations of the question, etc. (Sec.). To address this limitation, some recent methods (Li et al. 2023b) have attempted to artificially modify the format of generated answers to align with the reference answers. However, we argue this adjustment is a consequence of the flawed evaluation metric and should not influence modeling. Although human evaluation is the most reliable method for assessing generative models, it is costly and not scalable. Thus, there is a need to develop more robust automatic VQA metrics that serve as a proxy for human judgment.

A potential solution towards this issue would be to use soft evaluation metrics based on answer similarity (e.g.,

BERTScore (Zhang et al. 2020), Sentence-BERT (Reimers and Gurevych 2019)). While these metrics might be effective in matching synonyms and paraphrases, they fail when the compared texts have fine-grained yet major differences (e.g., “the man on the *left*” vs. “the man on the *right*”, “there *is a* dog” vs. “there *is no* dog”). We empirically evaluated the performance of such soft metrics for VQA, and found that their correlation with human judgement is even weaker than that of VQA Accuracy (Sec.).

Inspired by recent advances in using large language models (LLMs) to evaluate natural language generation (NLG) (Fu et al. 2023; Liu et al. 2023; Zheng et al. 2023), we explore the potential of leveraging LLMs as superior evaluators of answer quality in VQA. We believe that LLMs have the potential to capture human preference given their extensive training in modeling human language, and hence present a compelling choice as proxy for human judgment. By employing LLMs, we can harness the benefits of soft metrics while mitigating their limitations, resulting in a more robust evaluation framework.

To this end, we propose a novel automatic VQA evaluation metric, *LAVE* (LLM-Assisted VQA Evaluation), which leverages the in-context learning capabilities of instruction-tuned LLMs. In particular, we formulate VQA evaluation as an answer-rating task where the LLM is instructed to score the correctness of a candidate answer given the corresponding question and a set of reference answers. To evaluate the effectiveness of the proposed metric, we collect human judgments on the correctness of answers generated by several state-of-the-art VQA models across three popular VQA benchmarks. Our results demonstrate that LAVE correlates better with human judgment compared to existing metrics in diverse settings (Fig. 1). We also systematically categorize the failure modes of VQA Accuracy and show that LAVE is able to recover most missed correct candidate answers. In addition, we conduct ablation studies to assess the impact of each design choice on the performance of LAVE. In summary, our contributions are:

- We propose a novel metric for automatic VQA evaluation, LAVE, leveraging the in-context learning capabilities of instruction-tuned LLMs.
- We rigorously assess the effectiveness of LAVE by computing its correlation with human judgment, and show its robustness across various VQA models and benchmarks.
- We benchmark several strong baseline metrics in addition to VQA Accuracy, such as BERTScore or S-BERTScore, and show LAVE outperforms all of them.
- We systematically categorize the failure modes of VQA Accuracy and show LAVE fixes most of its pitfalls.
- We conduct thorough ablation experiments to measure the effect of each component of LAVE.

Related Work

Metrics for VQA VQA evaluation has received limited attention since the original VQA Accuracy metric was introduced by Antol et al. (2015). In a later study, Luo et al. (2021) propose enhancing the reference answers with alternative answer sets (AAS), focusing on the case where

only one reference answer per question is provided. More recently, Hu et al. (2022) devise a soft VQA Accuracy metric as part of their data filtering pipeline. We implement this metric as one of our baselines for comparison (Sec.).

Metrics for GenQA VQA and QA both involve answering questions related to a given context, either visual or textual. Generative QA evaluation faces similar challenges as VQA, relying primarily on metrics such as exact-match (EM) and F1 Score. Similarly to Luo et al. (2021), Si, Zhao, and Boyd-Graber (2021) also propose expanding reference answers with equivalent ones mined from knowledge bases. Lee et al. (2021) introduce a metric that weights answer tokens via keyphrase prediction. Chen et al. (2019) found that a straightforward application of BERTScore fails to provide stronger correlation with human judgements. Instead, other works (Risch et al. 2021; Bulian et al. 2022) train a semantic answer similarity metric based on BERT, showing improved correlation with human judgment. In contrast, we explore the capabilities of instruction-tuned LLMs in comparing candidate and reference answers.

Using LLMs as evaluators Recently, several works (Fu et al. 2023; Liu et al. 2023; Kamaloo et al. 2023; Li et al. 2023a; Zheng et al. 2023; Rajani et al. 2023) have explored the possibility of using LLMs (Flan-T5 (Chung et al. 2022), OPT (Zhang et al. 2022), GPT-X (Brown et al. 2020; OpenAI 2023)) to evaluate text generation for different tasks (e.g., summarization, dialogue generation, machine translation, QA, ...). Closer to our work, Zhou et al. (2023) propose using ChatGPT to automatically evaluate model outputs on a Likert scale. However, the quality of their metric remains uncertain as they do not provide any evidence of its alignment with human judgment. In this work, we aim to rigorously assess the effectiveness of LLMs in evaluating VQA by measuring their correlation with human judgment in diverse settings.

Analysis of VQA Accuracy Failure Modes

The motivation for developing a new VQA metric arises from the limitations of VQA Accuracy in handling open-ended model responses, which are not suitable for exact



Figure 2: Example of a VQA Accuracy failure mode from the *multiple answers* category (Tab. 1). Q: What are the sticks for? A: balance, pushing, skating, ...

Category	Definition	Examples	%
Multiple answers	Subjective, answers might focus on different aspects of the scene/activity.	Q: What are the sticks for? A: <i>balance, pushing, skating, ...</i> (Fig. 2)	34.25
Over- or under-specifying and verbosity	The candidate answer contains more/less details than the references or is more/less verbose.	Q: Where is the hydrant? A: <i>on the right, right</i> ; Q: What color are the flowers? A: <i>pink, pink and orange and red</i>	27.75
Synonym	Includes “almost-synonym” relation.	Q: What is the setting of this picture? A: <i>field, plains, grassland</i> ; Q: What is the sign telling you to do or not do? A: <i>no entry, do not enter</i>	21.0
Broad/bad question or generic response	Question is near impossible to answer or highly subjective; model avoids answering by being overly generic.	Q: How many sheep are there? A: <i>many</i> ; Q: What is the current condition of these animals? (image simply shows a baby elephant)	18.0
Incorrect	Human judgment is incorrect.	-	8.25
Same stem	Reference and candidate share the same stem (plural vs. singular or gerund); different formatting or whitespace.	Q: What are the people doing? A: <i>playing video games/game</i> ; Q: What shape is the building? A: <i>rectangular, rectangle</i> ; Q: What colors are in the surfer’s shirt? A: <i>blue and white, white and blue</i>	5.75
Hypernym	“Subcategory-of” relation.	Q: What are the people doing? A: <i>playing wii, playing video games</i> ; Q: What is in the blender? A: <i>vegetables, carrots</i>	5.0
Unknown issue	Model responds with “unknown”.	-	3.75
Ambiguous object	Phrase could refer to multiple objects in the image.	Q: What kind of sign is this? A: <i>billboard, street sign</i> (image shows multiple signs/billboards)	2.0

Table 1: Failure modes of the strict VQA Accuracy where correct responses are marked as incorrect. Model generated answers are marked in *italics* and reference answers underlined.

string matching. To understand the specific failure modes a new metric should address, we conducted a small study where we manually categorized 400 VQA examples. We looked at examples where VQA Accuracy is below 0.5 (at most 1 out of 10 reference answers matches with the model’s response), but at least 4 out of 5 humans rated the model’s response as correct. In other words: when are actually correct responses marked as incorrect the way current VQA systems are evaluated? We annotated 100 examples for each of four model-dataset pairs: (BLIP-2, VQAv2), (BLIP-2, VG-QA), (BLIP-2, OK-VQA), and (PromptCap, VQAv2). We focus on BLIP-2 and PromptCap since their generation is most open-ended.

Our initial set of failure modes is inspired from Luo et al. (2021), and manual inspection resulted in several additional categories. For clarity and conciseness, we decided to merge certain categories. Tab. 1 shows the consolidated nine categories with definitions, examples and frequencies.

We identified four prevalent failure modes: (1) **multiple answers**, (2) **over- or under-specification and verbosity**, (3) **synonyms** and (4) **broad/bad question or generic response**. We observe that certain question types naturally lead to various possible correct answers. For instance, many where-questions (e.g., “Where is the clock?”) can be answered using either absolute positioning or relative positioning to other objects. Other open-ended questions, such as asking what a person is doing or feeling, can be interpreted in multiple ways (e.g., “Why is she posing for picture?”). Luo et al. (2021) introduced the category *ambiguous object* when a phrase in the question could point

to several objects (e.g., “What color is the shirt?” when there are several shirts). However, our inspection showed only a few occurrences of it and we speculate it often also falls into the *multiple answers* category.

In summary, our analysis revealed that the open-ended nature of visual question answering can lead to multiple complex failure modes in VQA Accuracy.

Method

We present LAVE, an LLM-based evaluation framework to automatically assess the quality of answers generated by VQA models. Each VQA example comprises an image i , a question about the image q , and a set of reference answers R provided by human annotators. Given i and q , a VQA model f generates a candidate answer c , i.e., $c = f(i, q)$. Our goal is to automatically evaluate the quality of the generated answer c by comparing it with the references R . To enhance the evaluation process, we can additionally leverage the contextual information from the question q and the image i . We build a textual prompt using R , c , and optionally q and i (as an image description). This prompt is then fed to an LLM to generate a quality rating. The following sections describe the key design decisions underlying our approach.

Choosing a Large Language Model

It is crucial to choose an appropriate LLM as LAVE’s performance directly hinges on its capabilities. We pose VQA evaluation as a close-ended answer-verification task and adapt a frozen LLM through in-context learning. Hence, we opt for an instruction-tuned LLM, which has demonstrated

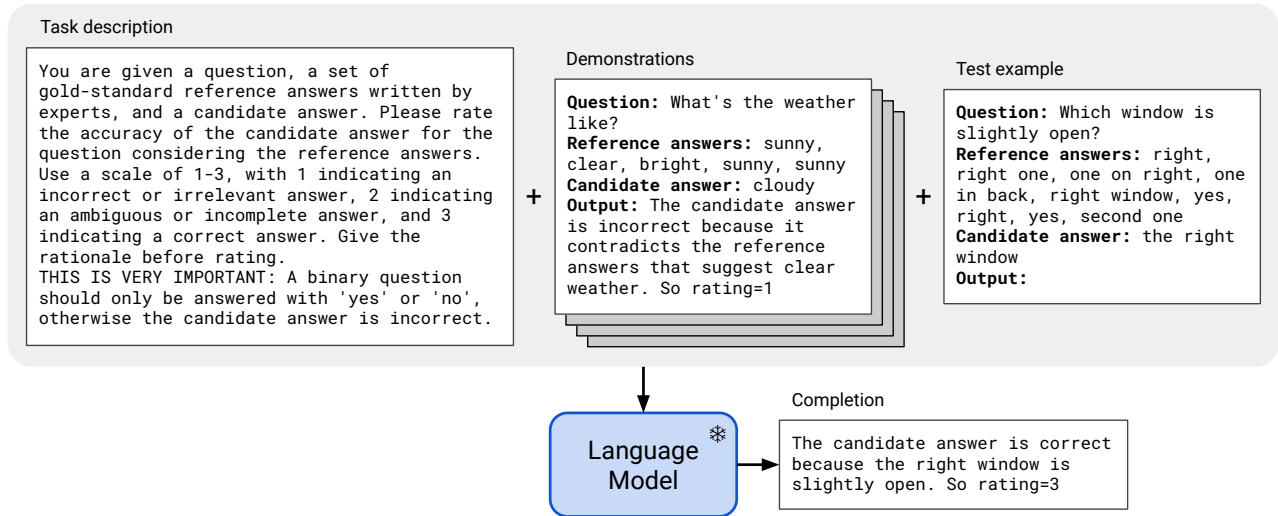


Figure 3: VQA evaluation with an LLM via in-context learning.

superior performance in transferring to new tasks with limited demonstrations (Wei et al. 2022a). Instruction-tuned LLMs are also more robust to prompt selection, and they can match the few-shot performance of much larger LLMs pretrained with self-supervised objectives. Considering all these factors, we first select the Flan-T5 (Chung et al. 2022) model family for our metric. In addition, Flan-T5 is finetuned on chain-of-thought (CoT) data, enabling it to provide reasoning for its answers.

To demonstrate LAVE’s robustness across different LLMs, we also consider Vicuna-v1.3 (Chiang et al. 2023) and GPT-3.5-Turbo (aka ChatGPT (OpenAI 2022)). We optimize our prompt for Flan-T5 (Sec.) and subsequently use the same prompt with the other LLMs. This opens the possibility of enhancing our metric without extra effort as stronger LLMs become available in the future.

Prompt for VQA Evaluation

We frame VQA evaluation as an answer-rating task amenable to in-context learning with LLMs. We adopt a rating scale ranging from 1 to 3 (as opposed to a binary rating) to account for ambiguous questions or incomplete answers. Our prompt (Fig. 3) comprises the typical components: task description, a few demonstrations of input/output, and the input for a test example. We draw inspiration from SNI (Wang et al. 2022) to structure our task description, as it is one of the main sources of training data for instruction-tuned LLMs. We also append “Give the rationale before rating.” to elicit a justification for the assigned rating, which improves explainability. Each demonstration consists of a question q , a set of reference answers R , the candidate answer c , the answer rating r , and an explanation e for the rating. We observed binary questions are particularly challenging to evaluate (App.), so we manually curate two sets of 8 demonstrations, one for binary questions and the other for general questions. We ensure demonstrations are diverse and cover various question types, num-

bers of reference answers, levels of agreement, candidate answer precision and verbosity, ratings, etc (refer to App. for the complete list). While these sets of demonstrations are designed to be comprehensive, users of our metric could also provide their own demonstrations to cover different cases. Additionally, to account for noise in the annotations, we filter out outlier reference answers that have a frequency lower than 25% of the maximum answer frequency. Finally, the test example only includes q , R and c , and the LLM is expected to generate an explanation e followed by a rating r . We found incorporating visual context from the image i into the prompt does not provide significant benefits (see Sec. and App. for more details).

Scoring Function

Given the LLM’s generated text for the test example, we extract the rating r from the last character (either 1, 2 or 3) and linearly map it to a score s in the range $[0, 1]$: $s = (r - 1)/2$. Inspired by Liu et al. (2023), we also explored the possibility of using the probabilities of output tokens to normalize the ratings and take their weighted sum as the final rating, but we did not observe any improvements in our task.

Experiments

Experimental Setup

VQA models and benchmarks We evaluate LAVE on answers generated by several VQA models across multiple VQA benchmarks. In particular, we consider two representative state-of-the-art VQA models: **BLIP-2 Flan-T5-XXL** (Li et al. 2023b) and **PromptCap GPT-3** (Hu et al. 2022). Our selection criteria were based on their public availability, their zero-shot VQA capability, and their architectural diversity. We also include **BLIP** (Li et al. 2022) finetuned on VQA_{v2} (BLIP_{VQA}) and VG-QA (BLIP_{VG}), which represents the finetuning-OOD paradigm. We use these VQA models to generate answers for three VQA

datasets: **VQAv2** (Goyal et al. 2017), **VG-QA** (Krishna et al. 2017) and **OK-VQA** (Marino et al. 2019). The selection of these datasets was driven by their diverse answer distributions. VQAv2 was prioritized due to its popularity as one of the most widely-used VQA benchmarks, providing 10 reference answers per question. VG-QA was chosen for its notably distinct answer distribution compared to VQAv2 (as shown by Agrawal et al. (2023)), and it provides a single reference answer per question. Lastly, OK-VQA was selected for its unique answer distribution, differing from both VQAv2 and VG-QA, as some of its questions require external knowledge to answer.

Baselines We evaluate LAVE against several strong baselines for VQA evaluation which involve comparing a candidate answer with a set of references. We consider the original **VQA Accuracy** (Antol et al. 2015), based on exact string matching; **soft VQA Accuracy** (Hu et al. 2022), which replaces exact-match by edit distance (CER); **ME-TEOR** (Banerjee and Lavie 2005), which uses unigram matching on surface form, stemmed form and meaning; **CIDEr** (Vedantam, Lawrence Zitnick, and Parikh 2015), which captures consensus among multiple references; **BERTScore** (Zhang et al. 2020), which calculates pairwise cosine similarity of contextualized token embeddings; and **S-BERTScore** (Reimers and Gurevych 2019), which measures cosine similarity of sentence embeddings. Both BERTScore and S-BERTScore compute similarity between pairs of sentences, so when there are multiple reference answers, the maximum score with the candidate is selected.

Implementation details We consider Flan-T5-XXL and Vicuna-v1.3-13B as open-source LLMs, and GPT-3.5-Turbo (gpt-3.5-turbo-0613) as a closed-source LLM. We leverage the HuggingFace Transformers’ (Wolf et al. 2020) implementation of Flan-T5 and LLaMA (for Vicuna), and use GPT-3.5-Turbo through OpenAI’s API¹. To make generation deterministic, we perform greedy decoding, or equivalently set the temperature to 0 in OpenAI’s API.

Collecting Human Judgments

We collected human judgments about answer quality using Amazon Mechanical Turk (MTurk) with the same web interface as Agrawal et al. (2023). Specifically, we collected judgments for answers generated by BLIP-2 on VQAv2, VG-QA and OK-VQA, PromptCap on VQAv2 and OK-VQA, BLIP_{VQA} on VG-QA and OK-VQA, and BLIP_{VG} on VQAv2 and OK-VQA (2450 questions each²). In total, our test set contains 22.1k questions. We additionally collected validation/development sets of human judgments for answers generated by BLIP-2 on VQAv2 and VG-QA, and BLIP on VQAv2 and VG-QA (1000 questions each). In total, our validation set contains 4k questions, which serve to guide our design choices. We emphasize that *PromptCap and OK-VQA are completely unseen during metric development* to show LAVE’s generality. Following (Agrawal et al.

2023), each answer was assessed by 5 annotators who were asked to provide a binary rating (correct/incorrect) based on the corresponding image and question. An important difference between the task posed to turkers and to the LLM is that the LLM is provided with a list of reference answers annotated by humans, while turkers are not. After filtering out low-quality annotations, we obtain an inter-annotator agreement measured by Krippendorff’s α of 62.0. Upon manual inspection, we observed that model responses (and even questions) are often ambiguous in nature, which would explain the relatively low inter-annotator agreement (see App. for more details). We derive a single “quality” score from the 5 binary ratings per answer as follows: 1.0 if at least 4 annotators rate the answer as correct, 0.5 if only 2 or 3 did so, and 0.0 otherwise. Considering partial scores is crucial to acknowledge the ambiguity inherent in certain questions, particularly when dealing with generative models which can produce technically accurate answers that may not align with the intended meaning of the question.

Correlation with Human Judgment

To evaluate LAVE, we measure its correlation with human judgment using two widely accepted rank correlation coefficients: Spearman’s ρ and Kendall’s τ (in the appendix). These provide a robust measure of the association between metrics, without assuming a linear relationship or a specific distribution of the data. Both coefficients range from -1 to 1 , with $-1/+1$ meaning perfect inverse/direct correlation.

Spearman correlations between VQA metrics (ours and baselines) and human judgment on every evaluated pair of VQA model and dataset are shown in Tab. 2 (see App. for additional results and observations). We verify statistical significance by bootstrapping with 5000 resamples and running a t-test with a significance level of 5% between pairs of correlations. For each setting, we bold the best results which are significantly better than the second-best result. The main observations are summarized as follows:

Overall, LAVE is significantly more aligned with human judgment than all the considered baselines, independently of the underlying LLM. Among the considered LLMs, we observe GPT-3.5 provides the highest overall correlation with human judgment. This is in line with recent LLM leaderboards (Zheng et al. 2023) which, at the time of writing, place GPT-3.5 (along GPT-4 (OpenAI 2023) and Claude (Anthropic 2023)) one step above any other LLM across a diverse set of benchmarks. However, we are excited to report that, on the VQA evaluation task, open-source LLMs such as Flan-T5 or Vicuna also outperform all baselines on average. Interestingly, despite being trained on user-shared conversations with ChatGPT, Vicuna falls behind Flan-T5 in our task.

LAVE generalizes to new VQA models and benchmarks. We did not use human judgments of answers produced by PromptCap (across all datasets) or to OK-VQA questions (from all models) to guide our prompt design (see Sec.). Still, our metric correlates better with human judgment than all baselines in these hold-out settings. This indicates our design choices are not overfitted to the particular settings used during metric development, and that LAVE ap-

¹<https://platform.openai.com/docs/api-reference>

²We initially collected human judgments on 2500 questions per model-dataset pair, but had to remove some to control data quality.

	BLIP-2		PromptCap		BLIP _{VG}		BLIP _{VQA}		Overall	
	VQAv2	VG-QA	OK-VQA	VQAv2	OK-VQA	VQAv2	OK-VQA	VG-QA		OK-VQA
Baselines										
VQA Acc.	71.54	41.19	48.65	65.82	48.24	84.85	70.38	41.90	68.84	60.13
Soft VQA Acc.	73.23	49.88	47.65	67.06	53.46	83.59	66.79	52.04	67.87	63.91
METEOR	64.75	48.70	50.97	57.74	51.76	83.45	71.42	51.65	68.44	58.68
CIDEr	69.55	47.78	53.23	63.26	49.81	85.07	71.08	46.50	70.25	63.88
BERTScore	50.61	11.73	38.62	41.14	42.73	72.14	59.42	15.88	60.51	31.47
S-BERTScore	60.44	42.10	47.84	47.44	47.11	77.65	68.04	47.61	65.61	56.61
Ours										
LAVE _{FT5}	71.19	59.94	59.85	64.18	58.67	71.67	66.03	54.50	63.87	64.99
LAVE _{Vicuna}	72.35	51.65	58.45	67.23	54.81	77.77	71.45	48.19	68.44	64.05
LAVE _{GPT-3.5}	74.25	60.19	61.47	71.99	57.39	83.63	69.97	58.47	67.57	68.91

Table 2: Spearman correlation (ρ) between VQA metrics and human judgment.

	BLIP-2		BLIP _{VG}	BLIP _{VQA}	Overall
	VQAv2	VG-QA	VQAv2	VG-QA	
LAVE _{FT5}	67.50	61.57	74.82	63.09	66.74
1-shot	55.43	61.04	59.79	57.45	59.07
4-shot	68.92	60.30	73.37	60.50	65.40
w/o rationale	58.67	63.87	68.01	65.36	65.57
w/o filter refs.	62.53	61.58	74.09	63.05	64.89
w/ caption	68.47	63.50	71.33	64.25	66.94

Table 3: Spearman correlation (ρ) between LAVE_{FT5} and human judgment when ablating for prompt design choices.

pears promising for evaluating answers generated by various models and across different datasets.

Questions from VQAv2 and OK-VQA answered by BLIP follow a different trend. In these settings, VQA Accuracy’s correlation with human judgment is considerably higher than for zero-shot VQA models, whereas LAVE has only a slightly higher correlation. We observe human score is much higher for BLIP-2 and PromptCap answers (0.7552 on average) compared to BLIP answers (0.5293 on average). Therefore, BLIP answers are more frequently incorrect or incomplete, which is expected as open-ended generative models are known to perform better on OOD data. The higher correlation for VQA Accuracy in these settings can be attributed to its efficacy in identifying incorrect candidate answers, while LLMs might label some as correct. For instance, GPT-3.5 labels “sink” as a correct answer to the question “What is this kind of sink called?”, or “refrigerator” as a correct answer to “What does this device generally do?”. Thus, the different trend in correlation with human judgment can be explained by a higher frequency of incorrect answers. This trend does not hold for VG-QA because LAVE outperforms the baselines when there is a single reference answer (see App.).

Ablation Studies

We compute correlation between LAVE_{FT5} and human judgment when ablating for different prompt design choices. The best overall configuration is used to compute correlation

on the test sets. Tab. 3 summarizes our ablation results.

Number of demonstrations Our results suggest a positive correlation between the number of demonstrations and LAVE_{FT5}’s effectiveness. As the number of demonstrations increases, the correlation with human judgment tends to improve. However, there is a tradeoff between number of demonstrations and computational overhead (and financial cost for GPT-3.5), so we tested up to 8 demonstrations.

Rationalization We measure the effect of asking the LLM to generate a rationale before rating candidate answers. Two trends arise when including rationalization: significantly improved performance on VQAv2 and slightly worse performance on VG-QA. We hypothesize a single reference answer (VG-QA) simplifies the answer-rating task, while having multiple reference answers (VQAv2) opens the door to discrepancies among annotators, leading to a more complex evaluation which can benefit from step-by-step reasoning.

Filtering of reference answers When using 10 reference answers (VQAv2), we observe that filtering out low-frequency answers consistently improves correlation, likely attributed to the reduction of noise in the reference answers. Note that this filtering has no effect when using a single reference answer (VG-QA).

Visual context We evaluate the effect of incorporating visual context into the prompt via an image description. Concretely, we use ground-truth captions from COCO and VG, modify the beginning of our task description as “You are given an image description, a question about the image, ...” and add “Image description: {caption}” to each example. Including visual context appears to be beneficial only in certain cases, especially when the dataset has a single reference answer (VG-QA). Notably, the overall correlation remains comparable to that observed without leveraging visual context. Moreover, in order to deploy our metric with visual context, we would need to add a captioning module to obtain the image description. Considering that visual context substantially increases the computational overhead due to image captioning and increased prompt length, we opted for excluding it from our final method.




	Image	Question	Reference answers	Candidate answer	VQA Acc.	LAVE _{GPT-3.5}	Rationale
(a)		What kind of ecological system is this?	water, lake, unanswerable, tropical, lush, cloudy, warm	aquatic	0.0	1.0	The candidate answer is correct because it is synonymous with 'water' and 'lake', which are mentioned in the reference answers.
(b)		What is ironic about this photo?	no water, boat, boat is out of water, boat not in water, boat is not in water, boat on dry land, boat is on dry land, boat on land, boat is on land	there is a boat on the street	0.0	1.0	The candidate answer is correct because it captures the irony of the situation where a boat is on the street instead of in the water.
(c)		What country's flag is shown?	usa, united states of america	american	0.0	1.0	The candidate answer is correct because 'american' is equivalent to 'usa' in this context.

Table 4: Selected questions from VQAv2 answered by BLIP-2, evaluated by VQA Accuracy and LAVE_{GPT-3.5}, along with the rationale for the answer rating. Duplicate reference answers have been omitted for conciseness.

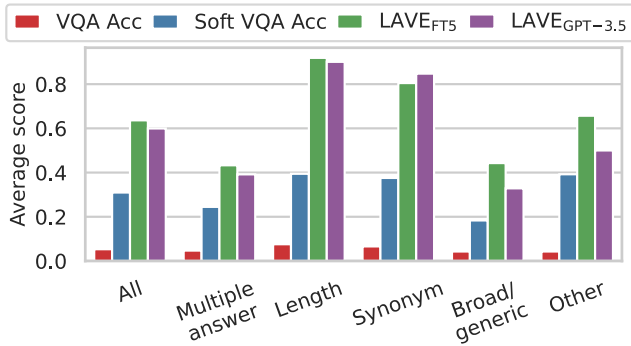


Figure 4: Average score of VQA evaluation metrics for cases where VQA Accuracy misses correct candidate answers, broken down by failure mode category.

Does LAVE Fix VQA Accuracy’s Failures?

Aside from having better overall correlation with human judgment, we would like to know how LAVE behaves in the failure modes of VQA Accuracy highlighted in Sec. . As a reminder, these are all cases where human annotators collectively labeled the candidate answer as correct (score of 1.0), while VQA Accuracy was below 0.5 (either 0.3 or 0.0). Therefore, we would expect our metric to give these candidate answers a score of 1.0 (excluding incorrect cases – 8.25%)³. Out of the 22.1k questions from our test sets, this is the case for 3601 questions (16.33%). For completeness, we found the reverse scenario, collective human score of 0.0 and VQA Accuracy above 0.5, occurs in 379 questions (1.72%); these are cases where the new human annotators disagree with the original annotations of the VQA datasets,

³Note that, in this setting, it is not possible to compute correlation with human judgment since it is constant (1.0).

indicating some noise in our collected human judgments.

Fig. 4 shows the average score of VQA Accuracy, Soft VQA Accuracy, LAVE_{FT5} and LAVE_{GPT-3.5} on the 400 manually-labeled examples analyzed in Sec. . We observe that LAVE is significantly more aligned with human judgment than both VQA Accuracy and Soft VQA Accuracy, especially when candidate answers are more verbose or they are a synonym of the reference answers. As previously mentioned, broad questions or which have multiple correct answers may be overly subjective, so it is harder for an LLM to determine whether the candidate answer is correct. It is interesting to see, however, that in these cases Flan-T5 generally performs better than GPT-3.5. In summary, this indicates that LAVE is able to recover a considerable fraction of correct candidate answers wrongly labeled as incorrect by VQA Accuracy.

Tab. 4 contains a few selected examples where LAVE_{GPT-3.5} fixes failures of VQA Accuracy. For instance, example (a) shows our metric is able to identify that the candidate answer is a synonym of several references, even though the form is different. Example (b) demonstrates our metric is robust to answers of diverse verbosity. In example (c), our metric is capable of identifying the candidate answer is equivalent to the references, even though they belong to different lexical categories.

Conclusions

We present LAVE, a new automatic VQA evaluation metric leveraging the in-context learning capabilities of instruction-tuned LLMs. Through a comprehensive study involving diverse VQA models and benchmarks, we demonstrate that LAVE is significantly more aligned with human judgment compared to existing metrics. We hope wide adoption of our metric will contribute to better estimating the progress of vision-language systems on the VQA task.

Ethical Statement

In this work, we propose a novel VQA evaluation metric leveraging the power of instruction-tuned LLMs. While this advancement has the potential to significantly improve the evaluation and development of VQA systems, it also raises several ethical and societal considerations that warrant careful attention.

First, while LAVE shows improved correlation with human judgment, we must acknowledge that the diversity and representativeness of the human annotators could influence the results. If the pool of annotators is not diverse, there may be biases in their judgments that could influence the performance of the proposed metric. We made a concerted effort to ensure that our pool of human annotators was as diverse as possible, but further research and mitigation strategies may be necessary to address this concern fully.

Second, the use of LLMs in any context brings up the issue of potential biases encoded in these models. As LLMs are typically trained on large-scale datasets scraped from the internet, they can inadvertently learn and perpetuate harmful biases present in those datasets. Such biases could result in discriminatory or otherwise unethical outcomes, so it is crucial to consider them when deploying or further developing LAVE. Future work should continue to investigate methods to identify and mitigate these biases.

Lastly, it is important to consider the broader impact of our research on society, particularly as it relates to the automation of tasks traditionally performed by humans. While improving VQA evaluation metrics could lead to more efficient and accurate systems, the potential displacement of jobs traditionally performed by humans could have significant societal impacts. It is essential to consider these potential consequences and to work towards solutions that leverage the benefits of AI while also considering the human factor.

Acknowledgments

We are grateful to Mila’s IDT team for their technical support with the computational infrastructure. The authors acknowledge the material support of NVIDIA in the form of computational resources. During this project, Aishwarya Agrawal was supported by the Canada CIFAR AI Chair award. We would also like to thank Samsung Electronics Co., Ltd. for funding this research.

References

Agrawal, A.; Kajic, I.; Bugliarello, E.; Davoodi, E.; Gergely, A.; Blunsom, P.; and Nematzadeh, A. 2023. Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization. In *Findings of the Association for Computational Linguistics: EACL 2023*, 1171–1196.

Anthropic. 2023. Introducing Claude.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bulian, J.; Buck, C.; Gajewski, W.; Boerschinger, B.; and Schuster, T. 2022. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Chen, A.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, 119–124.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Hu, Y.; Hua, H.; Yang, Z.; Shi, W.; Smith, N. A.; and Luo, J. 2022. PromptCap: Prompt-Guided Task-Aware Image Captioning. *arXiv preprint arXiv:2211.09699*.

Kamalloo, E.; Dziri, N.; Clarke, C. L.; and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. *arXiv preprint arXiv:2305.06984*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Lee, H.; Yoon, S.; Dernoncourt, F.; Kim, D. S.; Bui, T.; Shin, J.; and Jung, K. 2021. KPQA: A Metric for Generative Question Answering Using Keyphrase Weights. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2105–2115.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; and Liu, Z. 2023a. Mimic-it: Multi-modal in-context instruction tuning.

- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.
- Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Luo, M.; Sampat, S. K.; Tallman, R.; Zeng, Y.; Vancha, M.; Sajja, A.; and Baral, C. 2021. ‘Just because you are right, doesn’t mean I am wrong’: Overcoming a bottleneck in development and evaluation of Open-Ended VQA tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2766–2771.
- Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919.
- OpenAI. 2022. Introducing ChatGPT.
- OpenAI, R. 2023. GPT-4 technical report. *arXiv*, 2303–08774.
- Rajani, N.; Lambert, N.; Han, S.; Wang, J.; Nitski, O.; Beeching, E.; and Tunstall, L. 2023. Can foundation models label data like humans? *Hugging Face Blog*. <https://huggingface.co/blog/llm-leaderboard>.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.
- Risch, J.; Möller, T.; Gutsch, J.; and Pietsch, M. 2021. Semantic Answer Similarity for Evaluating Question Answering Models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, 149–157.
- Si, C.; Zhao, C.; and Boyd-Graber, J. 2021. What’s in a Name? Answer Equivalence For Open-Domain Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9623–9629.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Naik, A.; Ashok, A.; Dhanasekaran, A. S.; Arunkumar, A.; Stap, D.; et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5085–5109.
- Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E. H.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.
- Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.
- Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment.