# Selectively Answering Visual Questions

**Julian Martin Eisenschlos**[1,2], **Hernán Maina**[2,3], **Guido Ivetta**[2,3], **Luciana Benotti**[2,3]
Google DeepMind[1]
Universidad Nacional de Córdoba [2], CONICET, Argentina [3]
{julian.eisenschlos,hernan.maina,guidoivetta}@mi.unc.edu.ar
{luciana.benotti}@unc.edu.ar

## Abstract

Recently, large multi-modal models (LMMs) have emerged with the capacity to perform vision tasks such as captioning and visual question answering (VQA) with unprecedented accuracy. Applications such as helping the blind or visually impaired have a critical need for precise answers. It is specially important for models to be well *calibrated* and be able to quantify their uncertainty in order to *selectively* decide when to answer and when to abstain or ask for clarifications. We perform the first in-depth analysis of calibration methods and metrics for VQA with in-context learning LMMs. Studying VQA on two answerability benchmarks, we show that the likelihood score of visually grounded models is better calibrated than in their text-only counterparts for in-context learning, where sampling based methods are generally superior, but no clear winner arises. We propose AVG BLEU, a calibration score combining the benefits of both sampling and likelihood methods across modalities.

## 1 Introduction

Reliable Visual Question Answering (VQA) systems should provide a *confidence estimate* of their own predictions. This introspective skill allows users to decide when to trust or double check its outputs, or lets the system itself *selectively* decide when to gather more information in order to provide accurate responses.

While this problem has been studied extensively for classification models (Guo et al., 2017) and more recently also for text generation (Cole et al., 2023a), VQA systems have been under-explored. The combination of multiple input modalities can introduce different sources of uncertainty due to incorrectly framed or focused images. This is specially true when images are taken by people with no or limited vision, as is the case for the VizWiz-VQA dataset (Gurari et al., 2018). Furthermore the questions in VizWiz-VQA are spoken and therefore more conversational and contextual, which can introduce additional denotational uncertainty due to ambiguity when the question is under-specified because of shared common ground (Cole et al., 2023a). The multiple crowd-worker annotations in the dataset allow us also to identify confusing or unanswerable questions. It also allows us to see cases when even the unique answer to an unambiguous question can have multiple equivalent surface forms, or be expressed with different levels of specificity, for example in the month and year, or the full date for an expiration date.

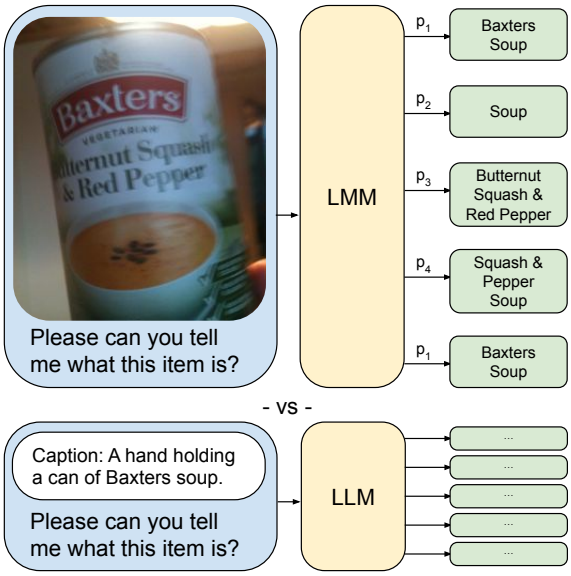


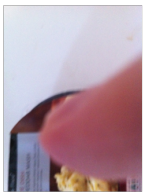

Figure 1: Sampled outputs from an LMM on a VizWiz-VQA example (Gurari et al., 2018) are used to measure *model calibration*. In this paper we contrast the LMM calibration results against LLMs that only sees the image caption. Sampling based methods struggle to measure uncertainty, motivating our proposed AVG BLEU as a *confidence estimate*.

Our goal is to study which methods from the literature can help calibrate state-of-the-art VQA systems, and build a *selective* VQA setup that can trust its confidence estimates when answering. We



*Findings of the Association for Computational Linguistics ACL 2024*, pages 4219–4229
August 11-16, 2024 ©2024 Association for Computational Linguistics

**Q**: When is the expiration date?



**Q**: What is this?

**Q**: Can you please tell me what the oven temperature is set to?

Figure 2: Unanswerable questions in VizWiz-VQA. Many are not answerable due to low quality images. Sometimes the intent of the question poser, and therefore the correct answer, cannot be inferred.

see this as a necessary step towards systems that can be used in real life scenarios. Figure 1 introduces three of the methods used in Cole et al. (2023a) as well as our proposed AVG BLEU (inspired in Wan et al. (2023)) on an example from the VizWiz dataset (Gurari et al., 2018).

Our contributions can be summarized as follows: **(a)** Through the use of several metrics, we study how the calibration methods from text QA perform when applied in two VQA answerability benchmarks and find that likelihood based scoring is better for LMMs than LLMs, without a clear winner. **(b)** We propose a new sampling based method that tackles limitations and improves all metrics significantly. Concretely, it improves coverage at $80\%$ accuracy by $5$ points for the best LMM and coverage at $70\%$ accuracy by $8$ points for the best LLM.

## 2 Related Work

Our paper studies VQA systems with unanswerable question through the VizWiz-VQA (Gurari et al., 2018) and UNK-VQA (Guo et al., 2023) datasets. We show some examples in Figure 2. Through the introduction of *unanswerable* responses from the annotators, Gurari et al. study a possible form of calibration as binary detection of such images. This setting is however limited in that a model (or person) can assess a question as *answerable* and be *uncertain* about whether the answer is a good one, as observed by Chen et al. (2023a).

On the topic of estimating uncertainty, several approaches appear in the literature. Collier et al. (2021) incorporate a latent trainable variable for the covariance matrix among classes. The use of additional binary classifiers— so-called *selectors*— was investigated for text QA (Kamath et al., 2020; Desai and Durrett, 2020; Jiang et al., 2021). This approach was extended to VQA by Whitehead et al. (2022) and Dancette et al. (2023). They focus on fine-tuned models and the use of selectors. We argue that, with the increasing availability and adoption of LMMs, it is important to study calibration in zero and few-shot scenarios, which pose unique challenges for evaluation (Maynez et al., 2023).

Cole et al. (2023a) study different sampling-based methods to calibrate LLMs in a zero-shot fashion. It focuses on entity based QA, where there is little differences in the surface form of equivalent answers. Due to the open-ended problem, the answers in VizWiz-VQA and UNK-VQA can vary, as illustrated in Figure 1. This variation needs to be considered when designing schemes to measure uncertainty when evaluating multiple samples.

## 3 Methods

A QA system is said to be well-calibrated when each prediction has a confidence score which can help assess how often it is correct. Formally, the paradigm of *selective QA* uses a scoring function $s$ that attaches to each QA pair $(q, a)$ a numeric score $s(q, a)$. The score can be compared with a threshold $\tau$ so that the system answers—aka *triggers*—when $s(q, a) > \tau$ or abstains otherwise. As shown in Figure 1, various scoring methods can be used. We present first the methods from Cole et al. (2023a) for text-only QA, and then our proposal.

**Likelihood** The likelihood-based calibration uses the output language model score $p(a|q)$ computed using the chain rule: $p(a|q) = \prod_{i=1}^{n} p(t_i|t_1, \cdots, t_{i-1}, q)$ where the $t_i$ are the tokens that form the answer $a$.

**Sampling Repetition** Based on Wilcox (1973)'s *Variation Ratio* we compute the frequency of the most sampled output divided by the total number of samples, which coincides with the probability of the mode of the empirical distribution. When more samples agree with each other, the answer can be considered to be more trustworthy.

**Sampling Diversity** Computed as $1 - \frac{\#\text{unique}}{\#\text{samples}}$, it is inversely proportional to the number of distinct samples and is zero if all samples are different.

**AVG BLEU** We propose an approach to grading the similarity among model answers that relies on averaging a measure of semantic similarity. Since there are multiple answers, we consider the average among all the possible pairs as an estimate of the diameter or dispersion of the full set. This is inspired by the measure proposed by Wan et al. (2023) for LLMs. Instead of ROUGE (Lin, 2004), which is not sensitive to

| | Large Multimodal Models (LMMs) | | | | | | | | | Large Language Models (LLMs) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | **LLaVA 13B** (61% ac@34% trig) | | | | **Flamingo 3B** (20.5% ac@71.7% trig) | | | | | **PaLM 2 Bison** (33.6% ac@78.8% trig) | | | | | **Falcon** (18.5% ac@87.0% trig) | | | | |
| Method | AUC | ECE↓ | C@70 | C@80 | AUC | ECE↓ | C@60 | C@70 | C@80 | AUC | ECE↓ | C@60 | C@70 | C@80 | AUC | ECE↓ | C@60 | C@70 | C@80 |
| AVG BLEU | **70.5** | 28.6 | **73.8** | **33.3** | 88.1 | 19.5 | 18.1 | 9.5 | **4.7** | 73.6 | **9.8** | **26.4** | **18.4** | **11.2** | 85.1 | **13.5** | **12.1** | **4.6** | 1.1 |
| Likelihood | 68.8 | 7.9 | 71.6 | 20.8 | **88.7** | 17.3 | **22.1** | **12.1** | 0.5 | 57.5 | 20.6 | 0.0 | 0.0 | 0.0 | 73.8 | 64.2 | 7.2 | 1.7 | **1.4** |
| Diversity | 70.4 | **6.8** | 67.2 | 8.7 | 77.6 | 12.4 | 11.4 | 7.2 | 3.7 | **75.0** | 39.9 | 10.1 | 10.1 | 10.1 | **86.2** | 31.3 | 11.8 | 0.0 | 0.0 |
| Repetitions | 69.3 | 9.9 | 73.2 | 28.4 | 76.7 | **5.4** | 7.9 | 7.9 | 4.4 | 72.7 | 31.8 | 21.6 | 10.1 | 10.1 | 84.0 | 34.0 | 11.8 | 0.0 | 0.0 |

Table 1: Calibration metrics on VizWiz-QA comparing LMMs on the left and LLMs on the right (by using gold image captions). We use 4-shots, except for LLaVa which only supports 0-shot. Best and second best values are **bolded** and underlined respectively. Likelihood lags behind sampling methods by more than 10 points for most metrics on LLMs but can surpass them in LMMs, although there is no clear winner. AVG BLEU combines both methods and performs above or comparable to the rest except for ECE, which can be fixed post-hoc by re-scaling.

| Model | **LLaVA 13B** (20% ac@65% trig) | | | | **Flamingo 3B** (12.5% ac@48.0% trig) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | ECE↓ | C@30 | C@40 | AUC | ECE↓ | C@20 | C@30 | C@40 |
| AVG BLEU | 61.0 | **27.3** | **26.4** | **1.3** | 71.9 | 9.9 | **50.3** | 17.4 | 2.4 |
| Likelihood | 60.1 | 49.5 | 22.1 | **1.3** | **72.4** | **3.8** | 48.6 | **18.4** | **5.2** |
| Diversity | 60.7 | 52.4 | 21.3 | 0.0 | 69.0 | 23.6 | 37.8 | 3.5 | 0.0 |
| Repetitions | **61.4** | 51.3 | 21.3 | 0.0 | 69.1 | 24.8 | 32.3 | 3.5 | 0.0 |

Table 2: Calibration metrics on UNK-QA. Best and second best values are **bolded** and underlined respectively. As is the case of VizWiz-VQA, likelihood is comparable or better than sampling methods. AVG BLEU performs above or comparable to the alternatives.

character n-grams, we use BLEU (Papineni et al., 2002) and compute the pairwise weighted average $\frac{1}{k}\sum_{i,j} p(a_i|q)\text{BLEU}(a_i, a_j)$ over the set of $k$ distinct predicted answers. Other similarity metrics are studied in Section 4.2. We fix the distance between a proper answer and *unanswerable* to 0.

## 4 Experiments

In our experiments, we use the validation split of VizWiz-VQA with 4k instances. Each question has up to 10 crowd-worker answers. Examples can be seen in Figure 2. We consider a question answerable if at least one crowd-worker annotated it as such. This corresponds to 75% of the questions.

We also include the validation set of UNK-VQA (Guo et al., 2023), consisting of 1K examples, synthetically constructed by modifying the images or text from VQA v2 (Goyal et al., 2017).

On the modeling side, we chose state-of-the-art LMMs LLaVA (Liu et al., 2023), Flamingo (Alayrac et al., 2022), and BLIP-2 (Li et al., 2023). To compare the calibration of LMMs with LLMs, we leverage the human written captions for VizWiz images provided by Gurari et al. (2020). We chose to use gold captions to control the additional errors and uncertainty which could arise from model written ones. Nevertheless we run a study in Section 4.1 using captions from PaLI-X (Chen et al., 2023b) and found that similar results hold. We evaluate state-of-the-art LLMs PaLM-2 (Anil et al., 2023) and Falcon (Almazrouei et al., 2023). We sample ten responses with a temperature of 0.7. We consider the model to *trigger* if the most likely answer (greedy) does not contain the sequence "unanswerable". The full prompts and BLIP-2 results can be seen in Appendix A. To simplify the analysis while staying consistent with the official metric defined in Antol et al. (2015), we consider a model answer correct if it matches at least one (instead of three) of the gold answers.

We evaluate the different scoring methods defined in the previous section with intrinsic and extrinsic metrics that we introduce below, over the set of examples where each model triggers a response.

**Expected Calibration Error (ECE)** Predictions are bucketed into ten same-sized bins, ranked by the confidence. We compute the mean absolute value of the distance between the average confidence score and the accuracy of predictions in each bin, averaged across all non empty bins. This intrinsic evaluation interprets a confidence score to represent a probability, so it computes the difference in the predicted probability of being correct from the observed probability of being correct. ECE is therefore noisy and sensitive to re-scaling.

**ROC-AUC** Area under the receiver operating characteristic curve measures the ability as a binary classifier for correct and incorrect predictions by integrating over the curve of the rates of true and false positives. It can be interpreted as the probability that the score of a correct output chosen at random is higher than that of an incorrect output.

**Coverage@Acc** While ECE and ROC-AUC assess absolute and relative calibration respectively, we want a metric closely aligned with selective QA. Therefore, we compute the triggering rate the model can achieve if forced to maintain a certain accuracy among the responses where it triggers. Formally, C@ Acc is the maximum coverage such that the accuracy on the C% of most-confident predictions is at least Acc %. For example, if

C@70 = 30, then the system is 70% or more accurate on the top 30% most-confident predictions. We take 100 coverage to mean not all of the examples, but where the model produced an answer.

We show the overall results for VizWiz in Table 1, including the baseline triggering rate and accuracy before incorporating calibration signals. The strength of the models is validated by comparing to fine-tuned models on the task, where a Flan-T5 (Base) (Chung et al., 2022) model trained on all the available captions is 74.2% accurate when it triggers, and a PaLI3-5 (Chen et al., 2023c) is 66.3% accurate when it triggers. Details about the fine-tuning experiments are explained in Appendix B. For LLMs, the findings of Cole et al. (2023a) hold, and likelihood is the worst calibration metric, repetition and diversity are superior to it. But, surprisingly, the picture changes for LMMs where we can see that repetition and diversity methods are not reliable. In contrast, AVG BLEU works well across the board, matching or improving the best metric in most cases. The same holds for UNK-VQA in Table 2. We omit C@Acc for accuracy below the model overall accuracy since it is 100.

We also evaluated model-based measures of semantic similarity in Section 4.2 beyond BLEU, but interestingly, we found that BLEU has strong performance while requiring negligible compute.

### 4.1 Human vs. model written captions

In order to restrict the noise and sources of uncertainty when evaluating LLMs for the VQA task, we opted to use gold captions. However, it can be argued that a more realistic setup would require the use of automated captions. With that goal, we evaluated the two LLMs on captions generated by a fine-tuned PaLI-X model (Chen et al., 2023b), which reported state-of-the-art results for the task when using an additional OCR module as input. Interestingly we observe improved results when using the model written captions (higher accuracy and triggering). Upon inspection of examples we see that the generated captions tend to be shorter and more concise, focusing on salient elements of the photo, which leads on average to better downstream answers for VizWiz where many questions are asking about salient objects (ex: *"What is this?"*).

As is the case for the manual captions, sampling based methods generally outperform likelihood by a large margin. The exception is on the ECE metric, which is expected given that it relies on the alignment between the scores and probabilities, can be fixed post-hoc, and is arguably the least practical one. We also observe that, AVG BLEU is able to be better or comparable to best result in every case.

| Model / Method | **PaLM 2 Bison** (45% ac@79% trig) AUC | ECE↓ | C@60 | C@70 | **Falcon** (32% ac@96% trig) AUC | ECE↓ | C@60 | C@70 |
|---|---|---|---|---|---|---|---|---|
| AVG BLEU | **75.4** | 12.1 | **58.3** | **37.8** | 74.8 | **7.9** | 22.1 | **1.5** |
| Likelihood | 63.1 | **11.0** | 28.5 | 4.1 | 67.9 | 21.1 | 2.4 | **1.5** |
| Diversity | 72.4 | 34.3 | 42.6 | 15.6 | **75.3** | 33.1 | **26.8** | 0.0 |
| Repetitions | 67.2 | 27.0 | 37.9 | 15.6 | 74.7 | 34.9 | **26.8** | 0.0 |

Table 3: Results on VizWiz-VQA using automated captions written by PaLI-X. Sampling based methods perform well except for ECE. AVG BLEU performs well across the board.

### 4.2 Replacing BLEU for dense similarity

While BLEU has the advantage of fast execution, it can fail to capture more nuanced forms of similarities among answers. We benchmarked two additional answer similarity metrics, BEM (Bulian et al., 2022) and BLEURT (Sellam et al., 2020). Perhaps surprisingly, the results over the various tasks show very little effect, as seen in Table 4. We speculate that the variability among 10 crowdworker answers diminishes the possible improvements, but it is possible that other VQA tasks could benefit from this approach making the additional compute cost worth the improved calibration.

| Method | AUC | ECE | C@60 |
|---|---|---|---|
| AVG BLEU | 88.1 | 19.5 | 18.1 |
| AVG BEM | 87.3 | 33.6 | 17.6 |
| AVG BLEURT | 88.6 | 18.8 | 22.3 |

Table 4: Comparison of calibration metrics for Flamingo predictions when replacing BLEU for trained similarity metrics. Both BLEURT and BEM give marginal gains for coverage at the cost of increased latency.

## 5 Analysis and Discussion

**How can we enhance the reliability of a VQA system?** Based on the results presented in this work and previous literature in the effect of calibration scores in user facing applications (Zhang et al., 2020) we can make the following recommendations when implementing a VQA system:

**(a)** Add a confidence score to the responses based on AVG BLEU that lets the users of VQA systems decide whether they can trust the system response or whether the question at hand is critical and needs a higher confidence in which case they can ask another person.

**(b)** Automated captioning and LLMs on those captions can be useful to answer visual questions

| Model / Method | Large Multimodal Model (LMM) Blip2 Flan T5-XL (39.3% acc @ 11.3% trig) AUC | ECE↓ | C@50 | C@60 | C@70 |
|---|---|---|---|---|---|
| AVG BLEU | 72.6 | **22.9** | **65.6** | 41.0 | **23.0** |
| Likelihood | 66.0 | 51.2 | 55.7 | 26.2 | **23.0** |
| Diversity | **78.7** | 29.5 | **65.6** | **49.2** | 0.0 |
| Repetitions | 75.8 | 32.6 | 57.4 | 45.9 | 0.0 |

Table 5: Calibration metrics on VizWiz-QA comparing for BLIP-2. As in the main results for other models, we see AVG BLEU performs above or comparable to the alternatives for most metrics and well above likelihood.

but likelihood is not a good calibration score in that case (Cole et al., 2023b).

**Why are LMMs better calibrated?** We speculate that this has to do with the inability for diversity and repetition metrics to capture similarity beyond text exact match and the grounding effect of the multi-modality acting as a regularizer and distributing the likelihood of possible answers in a more meaningful way. We leave further experiments in this direction to future work.

The language modality—e.g. gold captions—is a human generated signal with high information density and communicative intent (Rambow and Walker, 1994). Other non-communicative natural signal modalities have heavy redundancy, noise, and low information density, as observed in other multi-modal tasks (Wei et al., 2023). We can speculate that this produces a regularization effect—meaning the model training objective forces it to spread its bets among many answers. Grounding (Harnad, 1990) the question to sections of the image becomes particularly hard for images taken by visually impaired people in VizWizVQA due to occlusions or lack of framing, introducing an additional source of uncertainty (Chen et al., 2022).

**Can we measure accuracy better?** In the previous experiments, we perform exact-match (EM) criteria to classify whether the generated answer was correct or not, following the standard VQA accuracy evaluation (Antol et al., 2015).

This approach led to instances where the model, when presented with a question such as "What is my computer screen showing?" produced the response "A system restore". Despite the relevance of the generated answer, it was erroneously classified as incorrect due to the stringent acceptance criteria that required matching exactly the annotated answers, such as "system restore", "system restore message", "system restore pop up".

To address this limitation, we replicated the experiments incorporating three similarity techniques

| Method | EM | BLEU | Cos Sim | BEM (Bulian et al.) |
|---|---|---|---|---|
| Accuracy | 79% | 81% | 87% | 87% |
| AVG BLEU | **71** | **69** | 73 | 71 |
| Likelihood | 69 | 66 | 68 | 66 |
| Diversity | 70 | 67 | **75** | **72** |
| Repetitions | 69 | 66 | 70 | 69 |

Table 6: Comparison of ROC-AUC for confidence scores across different techniques for correct answer classification on LLaVa predictions. Even though accuracy is affected by the method of choice, AVG BLEU is better or comparable for all techniques.

for classifying correct answers, chosen based on their widespread adoption and use in comparable experiments (Risch et al., 2021). A similarity threshold was hand picked upon inspecting 20 instances. The results show variations in accuracy. Despite these differences, confidence metrics remained stable across the board, with variations of 8% at most, as shown in Table 6. Further details and experiments can be found in Appendix C.

**What errors do models make?** We manually inspected 272 errors made by the best two models on VizWiz-VQA presented in the previous section, one LLM and one LMM, with the goal of getting insights on their weaknesses. We find the following three most frequent type of errors. First, the LLM hallucinates more than the LMM and hallucinated answers in LLMs often trigger a response. Second, the gold caption of the LLM may not include the answer to the question. Third, the answer is true but not useful for visually impaired people. Appendix D illustrates each of them.

## 6 Conclusions

We studied scoring methods to assess the confidence in the predictions of a VQA system in the VizWiz and UNK-VQA datasets. We observed that, as shown previously, sampling based methods might present an advantage over likelihood based estimates in text models, but the picture changes for text-image models. Open ended question with non-entity answers present additional challenges for sampling methods, so the equivalence among possible sampled answers needs to be incorporated as well. Our AVG BLEU score is able to capture the spread of possible samples in a soft way and accomplishes the best results on the calibration evaluations we measured, combining the advantages of sampling and likelihood-based methods.

## Limitations

This study was conducted using several models over two datasets, which might limit the generalization of its conclusions. No other VQA datasets to our knowledge account for unanswerable questions in the same rich manner. For that reason, more analysis and datasets should be created to further research in this space and we hope our results serve as an initial step in that direction.

The use of manually curated captions may introduce a source of bias. Our analysis of error patterns reveals a substantial advantage for the LLMs due to its direct access to these gold captions. Addressing this limitation, future work could explore the feasibility and implications of working with automatically generated captions, allowing for a more realistic assessment in real-world scenarios.

## Ethical Considerations

In this paper we explore calibration methods for VQA models. This is an under-researched area which is particularly relevant for the visually impaired community for at least two reasons. First, quantifying the uncertainty in the day-to-day questions from this community can help them decide when to trust an automated VQA and when to ask another person. Second, our findings show that metrics in previous work have unstable performances for different kinds of models when evaluated over questions with high uncertainty such as those from visually impaired people.

The use of rating systems and visual question answering in the context of people with low or no vision carries risks since users of such systems cannot easily verify their results, and erroneous answers can lead to serious harm. In addition to accidental failures, models such as those studied can be attacked with the use of adversarial examples (Alzantot et al., 2018) by malicious actors with the aim of taking advantage of a vulnerable population.

For these reasons, before offering VQA systems for this population, it is essential to conduct detailed studies of the use cases and their possible failures, with emphasis on risk mitigation and respect of vulnerable populations (Le Ferrand et al., 2022). These studies should be designed collaboratively with members of the target group to minimize biases about how said system can be used.

## Acknowledgments

We would like to thank Francesco Piccinno, Srini Narayanan, Luciana Ferrer, and the anonymous reviewers for their time, constructive feedback, useful comments and suggestions about this work. This work used computational resources from both Google DeepMind and CCAD – Universidad Nacional de Córdoba[1], which are part of SNCAD – MinCyT, República Argentina.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 2890–2896. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 technical report.

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)*.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[1] https://ccad.unc.edu.ar

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19076–19085. IEEE.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023a. VQA therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15315–15325.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. Pali-x: On scaling up a multilingual vision and language model.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023c. PaLI-3 vision language models: Smaller, faster, stronger.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023a. Selectively answering ambiguous questions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023b. Selectively answering ambiguous questions.

Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. 2021. Correlated input-dependent label noise in large-scale image classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1551–1560.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. 2023. Improving selective visual question answering by learning from your peers. In *Computer Vision and Pattern Recognition*, pages 24049–24059, Los Alamitos, CA, USA. IEEE Computer Society.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 295–302. Association for Computational Linguistics.

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 *Computer Vision and Pattern Recognition*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Yanyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. 2023. Unk-vqa: A dataset and a probe into multi-modal large models' abstention ability.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Computer Vision and Pattern Recognition*.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020*, pages 417–434, Cham. Springer International.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an Australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4988–4998. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the Association for Computational Linguistics*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Owen Rambow and Marilyn Walker. 1994. The role of cognitive modeling in communicative intentions. In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4596–4604. PMLR.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, Hanjun Dai, Julian Martin Eisenschlos, Sercan O. Arik, and Tomas Pfister. 2023. Universal self-adaptive prompting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. 2023. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision – ECCV 2022*, page 148–166, Berlin, Heidelberg. Springer-Verlag.

Allen R. Wilcox. 1973. Indices of qualitative variation and political measurement. *The Western Political Quarterly*, 26(2):325–343.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA. Association for Computing Machinery.

## A Few-shot Experiments

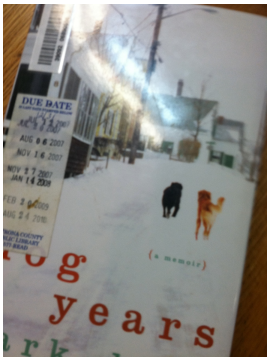
We show in Figure 3 and Figure 4 the 4-shot prompts used for LLMs and LMMs respectively, with representative examples chosen from the training set. The experiments were conducted in Tesla T4 GPUs with 16GB of VRAM and it took less than 4 hours for each model we executed. The results from BLIP-2 (Li et al., 2023) are shown in Table 5. We used the official models checkpoints released in https://hf.co. For PaLM2, we used the publicly available API[2] and the Flamingo predictions on the dataset were shared by the authors.

```
Read the descriptions of the following images and
    ↪ answer the questions using a single word or
    ↪ phrase. When the provided information is
    ↪ insufficient, respond with 'unanswerable'.

Descriptions:
- close up of a computer monitor that is powered on.
- A monitor has a message displayed on it.
- Pictured here is a screenshot that shows an error
    ↪ message from an app.
- Computer screen displaying an error saying the
    ↪ display driver is not supported by Zoom Text.
- a screenshot of someone's monitor that is having
    ↪ issues

Question: What does the arrow say?
Answer: unanswerable

Descriptions:
- a white paper showing an image of black and brown dog
- A library book with pictures of two dogs on the cover
    ↪  on a wooden table.
- A book with a black and a tan dog walking down a
    ↪ snowy street.
- The book cover shows two dogs in the snow
- A book cover title Dog Years with an image of a black
    ↪  and brown dog walking up the street, on the
    ↪ left side it has a due date sticker from a
    ↪ library

Question: What is the title of this book?
Answer: dog years

Descriptions:
- Quality issues are too severe to recognize visual
    ↪ content.
- A white object with elastic sides sitting on a wooden
    ↪  surface.
- A baby diaper keep in the table shown by the image.
- A opened up Diaper laying on a wooden surface.
- Clean maxi Pad over a dark wooden surface.

Question: What color is this
Answer: white

Descriptions:
- image is blur and hard to find what it is
- A woman in pink pajamas standing next to a white
    ↪ refrigerator in a kitchen.
- The back of a woman in pink standing at her counter
    ↪ and next to her white refrigerator
- A photo of a person wearing a pink shirt and pink
    ↪ flower pants in a kitchen next to a
    ↪ refrigerator.
- A person in red shirt and pink pajama pants stands in
    ↪  a kitchen near cabinets and counters.

Question: Is this a woman?
Answer: yes

Descriptions:
{descriptions}

Question: {question}
Answer:
```

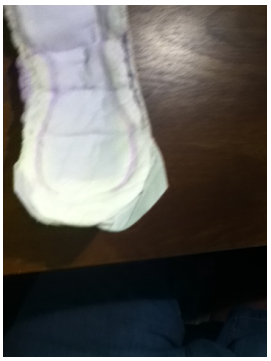
Figure 3: LLM 4-shot prompting for VizWiz-VQA using captions from Gurari et al. (2020).

```
Answer the question about the image using a single word
    ↪  or phrase. When the provided information is
    ↪ insufficient, respond with 'unanswerable'.
```

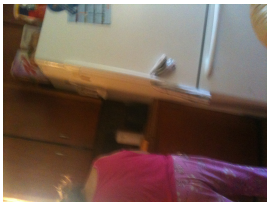

```
Question: What does the arrow say?
Answer: unanswerable
```




```
Question: What is the title of this book?
Answer: dog years
```



```
Question: What color is this
Answer: white
```



```
Question: Is this a woman?
Answer: yes

[IMAGE]

Question: {question}
Answer:
```

Figure 4: LMM 4-shot prompting for VizWiz-VQA.

## B Fine-tuning experiments

The Flan-T5 base model was trained using 3 NVIDIA GTX 1080Ti GPUs (GP102, 11 GiB

[2] https://cloud.google.com/vertex-ai

GDDR5) connected via PCIe 3.0 16x, with 32-bit floating-point precision. The fine-tuning process extended over 4k global steps until convergence was achieved, employing a batch size of 16. The learning rate schedule uses a linear warmup of 800 steps to 1e-5, followed by cosine decay to 0. Model optimization uses Adafactor (Shazeer and Stern, 2018). A sequence of 256 tokens was used for input encoding, while the output decoding utilized an 8-token sequence. The predictions for a fine-tuned PaLI-3 were shared by the authors and we refer to the paper for the training configuration. We show the results of the experiments in Table 7.

| Model | Acc | Trig |
|---|---|---|
| LLaVA (13B) | 61.2 | 34.0 |
| PaLM-2 (Bison) | 33.6 | 78.8 |
| PaLI-3 (5B) | 66.7 | 48.1 |
| Flan-T5 (Base) | 74.2 | 55.3 |

Table 7: Comparison of accuracy and triggering rate for in-context (top) vs fine-tuned (bottom) models. Few-shot LLMs abstain less than the other models.

## C Classifying correct answers

As discussed in Section 5, the experiments showcased in this paper used exact-match criteria to classify an answer as correct: only if the output generated by the model was exactly one of the accepted answers it counted as correct. In this section we describe the replication of these experiments, employing identical methodologies, introducing four distinct techniques for classifying correct answers. The selection of these techniques was based on their widespread adoption and their utilization in a comparable experiment outlined in (Risch et al., 2021). We perform this experiment for the LLaVA 13B and PaLM 2 Bison with similar results.

Each technique is briefly described as follows: Exact Match (EM) returns True only if the model's answer is exactly one of the accepted answers from the dataset. For Cosine Similarity we used *all-MiniLM-L6-v2* (Wang et al., 2020) as an embedding model due to its widespread use. BEM, as proposed in (Bulian et al., 2022), is a fine-tuned BERT model to classify answer equivalence using the SQuAD Dataset (Rajpurkar et al., 2016).

In Table 8 we compare the performance of the ECE method throughout the four tested techniques to classify correct answers. We can observe how both the accuracy and the value of ECE change, however the best performant method remains constant. These findings suggest that the results presented in this paper are robust across various answer classification techniques.

| Method | EM | BLEU | Cos Sim | BEM |
|---|---|---|---|---|
| Accuracy | 79% | 81% | 87% | 87% |
| AVG BLEU | 28.6 | 35.42 | 50.57 | 50.18 |
| Likelihood | 7.90 | 10.95 | 24.06 | 22.97 |
| Diversity | **6.80** | **6.17** | **13.39** | **12.30** |
| Repetitions | 9.9 | 11.20 | 20.49 | 19.62 |

Table 8: Comparison of ECE metric for confidence scores across different techniques for correct answer classification in LLaVa 13B. Accuracy and ECE metric are affected by the method of choice, nonetheless the best performant method (highlighted in bold) remains constant.

## D Insights from error analysis

In examining 272 errors from the top-performing language models LLaVA 13B which is an LMM and PaLM 2 Bison which is a LLM, we aimed to identify weaknesses specific to the needs of the vulnerable population. Three recurrent errors were observed which accounted for most errors. Firstly, the LLM tends to hallucinate more than the LMM, with these hallucinated responses influencing the final output. For example Figure 5 illustrates how PaLM 2 Bison hallucinates multiple serial numbers that appear neither in the image nor in the captions. This is less frequent for LLaVA 13B which refrains from answering in most of these cases.


Figure 5: Question *"Alright see if you can see the ORCA serial number now."*. The top 3 answers by PaLM2 Bison with their likelihood are ('31631036', -1.32), ('468563995', -1.41), ('4667926374', -1.41). They are neither in the image nor in the gold captions. The serial number is not in the image. Correctly, LLaVA 13B does not trigger a response in this case considering the question unanswerable.

Secondly, Figure 6 shows an example of a question that is not possible to answer with the information in the gold captions. This is to be expected a caption cannot include all the information present

in an image. This is a limitation of the methodology of using captions instead of images for VQA.

Lastly, the third kind of frequent error observed are answers that are factually correct, but are not useful for individuals with visual impairments. A example of this kind of error is shown in Figure 7.

These findings shed light on critical areas for improvement in models designed for this particular user group, showing that well calibrated systems are important in this domain.

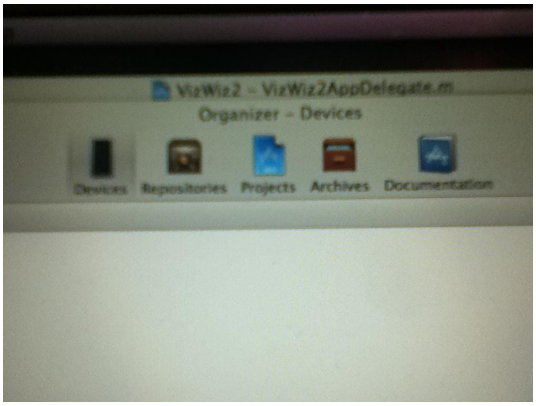



Figure 6: Question *"What action is currently selected?"*. The top 3 answers by PaLM2 Bison with their likelihood are ('projects', -0.69), ('archives', -0.76), ('devices', -1.13). The correct answer is devices but this information is not in the gold captions, captions can never be complete. The captions are 'The computer digital monitor screen with some websites open to', 'A quick link bar has appeared on an Apple computer screen.', 'Computer menu bar with options: Devices, Repositories, Projects, Archives, and Documentation.', 'some type of MacBook or laptop device u CNA use', 'A VizWiz app that has an organizer and lists projects and archives.'.

Figure 7: Question *"Which one is the blue one?"*.The top 3 answers by PaLM2 Bison with their likelihood are ('blue', -0.47), ('second', -0.63), ('right', -0.82). The first answer is true but useless for a visually impaired person. The top 3 answers by LLaVA 13B Bison are correct. They are ('right', -0.08), ('right', -0.08), ('right', -0.08). person.