Evaluating Modern Vision Language Model Zero-shot Performance on the TQA dataset

Team Name: ZeroShot

Jashwin Acharya, Nick Schnabel and Ahmed Shahkhan

1 Introduction

1.1 Motivation

Visual Question Answering (VQA) is a field that has witnessed many advancements in the past few years. Nowadays, there are many pre-trained Vision Language models available on websites like Hugging Face that allow students, researchers and engineers to leverage the power of these models on real world tasks such as VQA, Text Summarization, Dialogue Generation and many other utilities. One area of research that is less explored is the Zero-shot evaluation of modern Vision Language model capabilities on the Textbook Question Answering Dataset which contains "1,076 lessons and 26,260 multi-modal questions, taken from middle school science curricula" (Kembhavi et al., 2017). We believe it is important to understand the multi-modal generalization capabilities of these models in educational domains so that they can hopefully be used in the future for utilities such as automated grading, intelligent tutoring systems, and academic research.

1.2 Related Work

Kembhavi et al.'s paper from 2017 introduced the task of Multi-Modal Machine Comprehension in the educational domain which involves reading a multi-modal question and context, and producing an answer which could also be multi-modal in nature (Kembhavi et al., 2017). The authors evaluated 4 models on the TQA dataset: A Text Only model; two Text + Diagrams models; and a Machine comprehension model called BiDAF (Kembhavi et al. 2017) that stands for "Bidirectional Attention Flow for Machine Comprehension". The results of their experiment indicated very poor performance on the TQA dataset with the models having less than 35% accuracy on MCQ and True/False questions (Kembhavi et al. 2017).

While the original paper for the BLIP – short for "Bootstrapping Language-Image Pre-training" (Li et al., 2023) – model from 2022 did not evaluate Zero-shot performance on a VQA task, the authors of BLIP-2 proved that with less trainable parameters, they were able to achieve state-of-the-art results on the VQA task, scoring 65% VQA accuracy compared to the 56.3% accuracy of Flamningo despite having 54x fewer trainable parameters (Li et al., 2023). The prompt format used by the authors for the VQA task was "Question: Answer:" (Li et al., 2023).

The latest version of GPT-4 allows users to provide text in conjunction with images as prompts (also called GPT-4V (vision)). More recently, Wu et al., tested the zero-shot capability of GPT-4 using the ScienceQA dataset and their zero-shot analysis indicated a high accuracy of 85% (Wu et al., 2023), hinting at the generalization capabilities of GPT-4 with regards to topics from the educational domain. The authors also provided a couple examples of making GPT-4 guess the correct option for multiple choice questions about images they fed into GPT-4.

Models like CLIP in the past have shown great results in a zero-shot setting for a VQA task, and have managed to boost the accuracy in a few-shot setting. A paper from 2022 describes the performance of CLIP in a zero-shot setting which achieves an approximate accuracy of 71% on a VQAv2 validation set (Song et al., 2022), while fine-tuning in addition to parameter optimizations yielded slightly higher accuracies in the few-shot setting (Song et al., 2022).

2 Approach

2.1 Problem Statement and Hypotheses

Kembhavi et al.'s paper from 2017 highlighted the low accuracies achieved for the 4 baseline Text Only and Text + Image models. Since numerous

new models such as BLIP2, GPT 4, and LLaVA-7B have emerged since then, our research question is as follows:

Can modern state-of-the-art Vision Language models achieve better accuracies on the TQA dataset in a zero-shot setting?

To further investigate our research question, we derive three hypotheses, which we will answer by applying our model evaluation. The hypotheses are as follows:

- **Hypothesis I:** We believe that the Zero-shot performance of GPT-4, BLIP-2 and LLaVA-7B will be better than the original models from 2017 on MCQs that have images associated with them.
- **Hypothesis II:** We believe GPT-4 should provide better results on True/False questions and Non-Diagram MCQs than the original models from 2017 in a Zero-shot setting.
- **Hypothesis III:** We believe that few-shot learning could improve our accuracy compared to the Zero-shot setting.

2.2 Preparing the Data

We downloaded the TQA dataset from the Allen Institute for AI website and the data can be found here. The dataset is further divided into a train, test and validation set, but for the purpose of our zero-shot analysis we only focused on the test set as we wanted to compare our accuracies against the 2017 models evaluated on the TQA test set. The test data was in a JSON format which had to be converted into a regular CSV file so that it could be easily processed using the pandas library in python for our analysis. We wrote a python script that parsed the test JSON file and created three CSV files based on the three main sub-types of questions available:

- True/False questions where no image is provided and only a statement is given. The True/False CSV file contained 911 questions, answer choices and the correct answer for each question.
- Regular MCQs where no image is provided and only a question and multiple choice options are provided. The Non-diagram MCQ CSV file contained 1293 questions, multiple choice options, as well as the correct answer choices for each question.
- Diagram MCQs where an image is provided along with a question about the image whose answer is one of four options. The dataset for Diagram MCQs (3284 rows) contained details such as the path of the image, questions, multiple choice answers, as well as the right answers for each question.

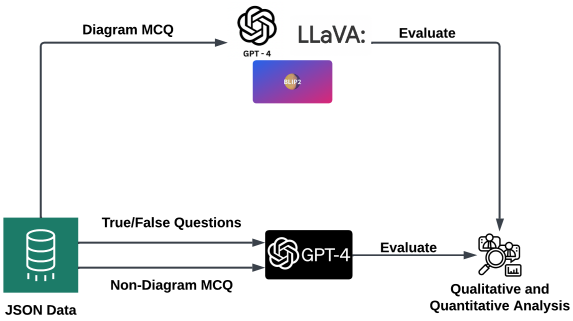

Figure 1: Project Workflow

Figure 1 shows how we extract the aforementioned three sub-types of questions from the test JSON file, feed the questions and answer choices into the relevant Vision Language Models (VLMs), and subsequently perform qualitative and quantitative analysis on the model results. Since BLIP-2 and LLaVA-7B require both visual and textual information to make inferences, we opted to only test the performance of GPT-4 on True/False and Regular MCQs as there were no images associated with these types of questions, and GPT-4 is capable of providing results based on just textual data. For regular Diagram MCQs, we test the performance of all 3 models.

2.3 Prompt Engineering

An important note to make is that the choice of the prompt has a great impact on the quality of answer produced by modern Vision Language Models.

2.3.1 Prompting with GPT-4

We used the following prompt structure for True/False questions when querying GPT-4:

Is this statement True or False: <Question>. Only tell me if its True or False. An explanation is not required.

We used the following prompt structure for both Diagram MCQs and Non-diagram MCQ questions when querying GPT-4 or GPT-4V:

Choose only one option below as the answer for the following question. An explanation is not needed.

Question: <Question>

<Answer Choice 1>
<Answer Choice 2>
<Answer Choice 3>
<Answer Choice 4>

We perform API requests by supplying our zero-shot prompts to the *Completions* endpoint provided by OpenAI's API and perform inferences using the *GPT-4 model* for True/False and Non-Diagram MCQs. For Diagram MCQs, we make inferences using the *GPT-4-Vision-Preview model* since we also pass it an image as part of the API request.

It's important to specify in the prompt that an explanation is not required since, initially, GPT-4 was providing us long answers to our questions which wasn't required in our case as we simply wanted to know what the correct answer choice was or if the statement was True or False for performing easy comparisons with the actual answers in our test CSV files.

2.3.2 Prompting with BLIP-2

We followed the instructions on the LAVIS GitHub page for setting up our environment for performing a VQA task with BLIP-2. The following prompt structure was used for Image MCQs when prompting BLIP-2:

Question: <Question> Choose only one option.

a. <Answer Choice 1>
b. <Answer Choice 2>
c. <Answer Choice 3>
d. <Answer Choice 4>

One important change to note here compared to the GPT-4 prompt is the addition of option labels such as **a, b, c and d**. When we didn't include those labels initially, BLIP-2 had a hard time understanding how to choose the answer choices. After adding those option labels, BLIP-2 was able to understand which options were available to choose an answer from and was able to provide inferences without any issues.

2.3.3 Prompting with LLaVA

We followed the instructions on the LLaVA GitHub page for setting up the LLaVA-7B model to perform inferences for a VQA task. We were unable to use the LLaVA-13B parameter model because of lack of GPU RAM; hence, we settled on the LLaVA-7B model. LLaVA uses the exact same prompt as GPT-4 for Diagram MCQs and subsection 2.3.1 can be referred to for details on the prompt structure.

2.4 Augmenting BLIP-2 and LLaVA prompts

Once we were able to generate inferences for Diagram MCQs using BLIP-2 and LLaVA-7B, we thought about adding image descriptions to our prompts to see if it could help boost accuracies or not. Since the TQA test set doesn't contain short image descriptions, our idea here was to find a subset of images ($\tilde{4}00$) that BLIP-2 and LLaVA-7B both struggled with and that GPT-4V had at least 80% success rate on and generate image descriptions using the GPT-4V. Owing to GPT-4V's high accuracy on those specific images, we thought it would be a viable idea to leverage its ability to generate rich image descriptions as an added context for BLIP-2 and LLaVA-7B prompts. The reason why we weren't able to generate image descriptions for all 3284 images is because GPT-4V has an API limit of 100 requests per day for the *GPT-4-Vision-Preview* model. Hence we decided to perform this evaluation only on a subset of examples as described earlier. For generating image descriptions using GPT-4V, we use a simple prompt such as ***"Give a brief 3-4 line description of this image."*** along with the image. The revised prompt structure for LLaVA-7B for making inferences on Diagram MCQs is below:

Image Description: <GPT-4 Generated Description>

Choose only one option below as the answer for the following question. An explanation is not needed.
Question: <Question>

<Answer Choice 1>

<Answer Choice 2>
<Answer Choice 3>
<Answer Choice 4>

The revised prompt structure for BLIP-2 for making inferences on Diagram MCQs is below:

Image Description: <GPT-4 Generated Description>

Question: <Question> Choose only one option.

a. <Answer Choice 1>
b. <Answer Choice 2>
c. <Answer Choice 3>
d. <Answer Choice 4>

3 Challenges

3.1 Setting up BLIP-2 and LLaVA

It was a major challenge trying to get the BLIP-2 and LLaVA-7B parameter models to work on our personal computers, so we opted to use MSI resources to run test set inferences on those models. Regardless of using a 48GB A40 GPU, we weren't able to use the bigger 13B parameter model for LLaVA as it requires around 55GBs of RAM to load the model weights. It also took almost a day to get all of our BLIP-2 and LLaVA predictions, which was another reason we opted to use a small subset of images for our approach mentioned in subsection 2.4.

3.2 Limitations with GPT-4

OpenAI's text completion API is generally quite reliable, but we randomly kept running into server issues when making API requests for our True/False, Non-diagram and Diagram MCQ prompts. The biggest limitation which we also mentioned in subsection 2.4 is that the GPT-4-Vision-Preview model only allows 100 requests per day, and since we had around 3284 images in our Diagram MCQ dataset, we had to split the work between all 3 team members in order to get all our predictions generated on time.

4 Experimentation Results and Analysis

4.1 True/False Question and Non-diagram MCQ Performance

Figures 2 and 3 show GPT-4's accuracy on True/False questions and Non-diagram MCQs

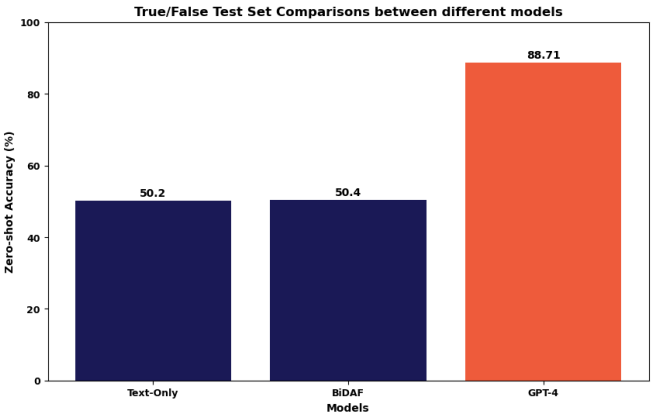

Figure 2: True/False Questions Accuracy

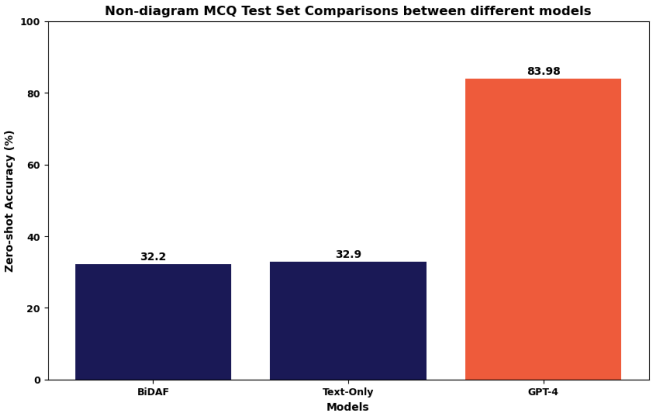

Figure 3: Non-diagram MCQs Accuracy

compared to Kembhavi et al.'s accuracies on their Text-Only and BiDAF models from 2017. As we can see, GPT-4 achieves a high accuracy of 88.71% and 83.98% on the test data for True/False questions and Non-diagram MCQs respectively. To the best of our knowledge, we don't think that GPT-4 has ever been trained on this dataset earlier, and the main reason for GPT-4's high performance on the test set for these questions is simply because of the large corpus of data that GPT-4 has already been trained on which must definitely contain many scientific questions similar to ones in the TQA dataset. Our results show that GPT-4 is able to generalize quite well on these questions without much added context, thus showcasing its powerful zero-shot performance as well as proving Hypothesis II defined earlier in subsection 2.1. It is difficult to exactly judge why GPT-4 gets a certain number of questions wrong for non-diagram MCQs and True/False Questions. The questions that GPT-4 gets right or wrong are quite similar and this indicates that while GPT-4's contextual understanding is strong, it isn't always perfect.

Hallucination is a known issue in many LLMs and it's entirely possible that GPT-4 simply produced the wrong answer because it had the wrong understanding of what was being asked in the question. The performance is still quite high and we are optimistic that few-shot or CoT strategies would have yielded even higher accuracies on these subtypes of questions.

4.2 Diagram MCQ performance without Image Descriptions

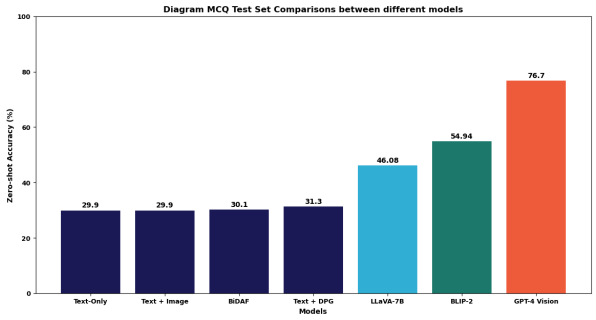

Figure 4: Diagram MCQ Accuracy without Image Descriptions

Figure 4 shows BLIP-2, LLaVA-7B and GPT-4V's performance against all the 2017 models on Diagram MCQs. All three aforementioned models beat the original paper's model in a zero-shot setting, further proving the power of modern Vision Language Models, and also proving Hypothesis I defined by us in subsection 2.1. GPT-4V performs the best with a 76.7% accuracy on the test data, followed by BLIP-2 at 54.94% and LLaVA-7B model at 46.08% accuracy. It's quite interesting to see that BLIP-2 with 2.7B parameters is able to generalize on the Diagram MCQs much better than the bigger LLaVA model, with GPT-4V showing the best generalization performance.

4.3 Diagram MCQ performance with Image Descriptions for BLIP-2 and LLaVA

As we can see in Figure 5, augmenting our zero-shot prompts for BLIP-2 and LLaVA-7B with GPT-4V generated image descriptions (as described in subsection 2.4) does allow BLIP-2 and LLaVA to correctly guess a small subset of answers correctly compared to our original prompts where image descriptions were not provided. The LLaVA-7B model, however, still has a low accuracy of 27.29% on our small subset of 403 images, whereas BLIP-2 continues to show stronger generalization performance with an accuracy of 52.6% which means that it was able to guess over half of the previous subset of miss-classified questions now correctly with the added image descriptions.

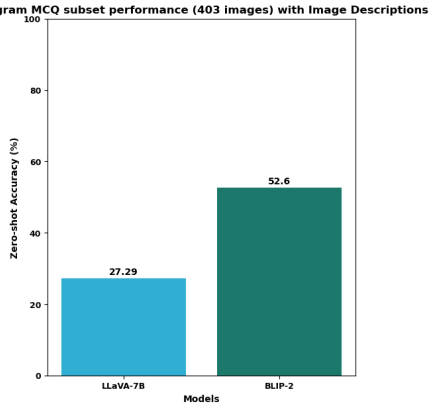

Figure 5: Diagram MCQ subset performance (403 rows) with Image Descriptions

4.4 Qualitative Analysis of modern VLMs

Apart from just providing a quantitative assessment of the power of modern Vision Language Models, we believe that a qualitative analyses of the model results is also apropos to our current case study of the TQA dataset. As part of our qualitative analysis, we aim to focus on what types of images and questions do BLIP-2, LLaVA and GPT-4 mostly struggle with and see if there's a significant failure pattern on certain types of images/questions over others.

4.4.1 Qualitative Analysis of BLIP-2 and LLaVA predictions on Diagram MCQs (without Image Descriptions)

Our qualitative analysis for these models indicated that performance tends to be a bit random with both models at times struggling with questions about simple images but being able to answer questions on complex images, and vice versa.

Figure 6 and 7 are two slightly complex images where BLIP-2 and LLaVA-7B show poor performance on the former but are able to guess half the answers correctly for the latter image which has far more labels in comparison. The type of questions doesn't seem to matter in this case because both BLIP-2 and LLaVA-7B are sometimes able to answer general questions correctly about the images such as identifying labels and their functions, and other times they provide an incorrect answer.

Figures 8 and 9 further show how BLIP-2 is able

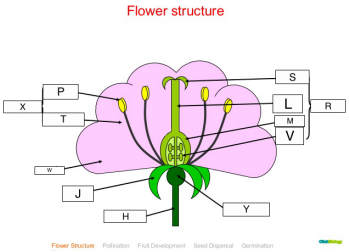

Figure 6: 5/6 questions answered incorrectly by BLIP-2 and LLaVA-7B

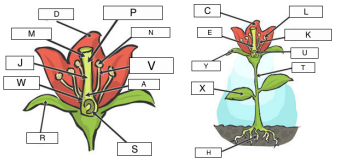

Figure 7: 3/6 questions answered correctly by BLIP-2 and LLaVA-7B

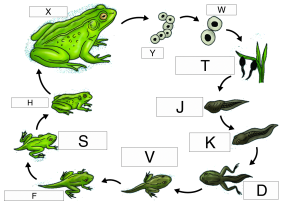

Figure 8: 3/5 questions answered correctly by BLIP-2

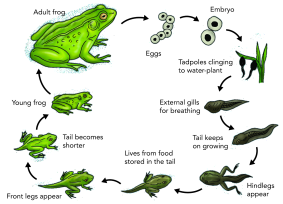

Figure 9: 5/7 questions answered incorrectly by BLIP-2 and LLaVA

to answer majority of the questions correctly for an unlabeled frog life cycle image as shown in the former figure, but has a hard time answering majority of questions about the same image but with labels even when the questions for both images are very similar in nature. This shows that BLIP-2's contextual understanding is a bit flawed when there is a lot of information in an image.

Figure 10 shows how BLIP-2 is unable to answer a single question correctly for a simple chicken life cycle image, but is able to answer most questions correctly for Figure 8 which also doesn't have labels, but is a more complex image than Figure 10 which continues to show the randomness of BLIP-2's performance when it comes to image complexity.

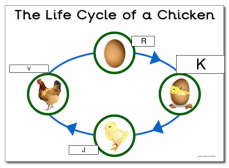

Figure 10: 5/5 questions answered incorrectly by BLIP-2

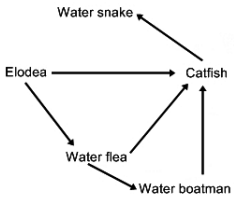

Figure 11: 8/8 questions answered incorrectly by LLaVA-7B

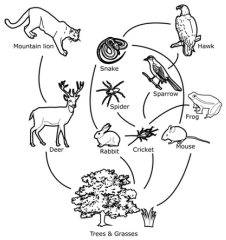

Figure 12: 7/8 questions answered correctly by LLaVA-7B

As can be seen in Figures 11 and 12, LLaVA-7B is unable to answer any question correctly for the simple food chain shown in Figure 11, but is able to answer 7/8 questions correctly for the far more complicated food chain in Figure 12. The questions are also similar for both images and are generally about what animal is a producer, which animal consumes another animal and so on.

The limitations of a zero-shot approach are quite apparent in our BLIP-2 and LLaVA-7B evaluation on the TQA dataset and we believe that in order to address these limitations, it would be ideal to fine-tune the models on the train set also provided as part of the TQA dataset. Fine-tuning almost always leads to better test set performance. We also believe that employing a few-shot or chain-of-thought approach can be helpful and

dive more into these approaches as part of Section 5.

4.4.2 Qualitative Analysis of BLIP-2 and LLaVA-7B predictions on Diagram MCQs (with Image Descriptions)

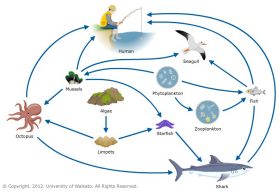

Figure 13: LLaVA-7B answered 3/5 questions incorrectly with image descriptions and 4/5 without.

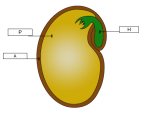

Figure 14: LLaVA-7B answered 3/4 questions incorrectly with image descriptions and 4/4 without.

As we can see in Figures 13 and 14, LLaVA-7B is not able to successfully leverage the provided image descriptions to improve its performance on most simple and complex images. It continues to struggle with the images in the figures as it did previously too which shows LLaVA-7B's lack of contextual understanding. Our analysis found that it's only able to answer around 111 questions correctly out of 403, leading us to believe that adding GPT-4 image descriptions provided only a little benefit, and maybe human-expert generated image descriptions might prove to be more helpful. It does provide slightly better performance on some images, but most of those images only have 1 or 2 questions associated with them. Thus, it is difficult to judge whether LLaVA is leveraging the image descriptions correctly or not.

Figures 15 and 16 show that BLIP-2 has a far better contextual understanding strength compared to LLaVA as it is able to achieve a positive success rate on some images that it previously struggled with because of the presence of image descriptions. However, it still struggles with some images such as the leaves in Figure 18 and the life cycle shown in Figure 17 which indicates to us that the lack of labels is causing BLIP-2 to still perform poorly on these images. Since we are only using a subset of images for our image description analysis, it would be interesting to see how much BLIP-2's performance could be boosted on

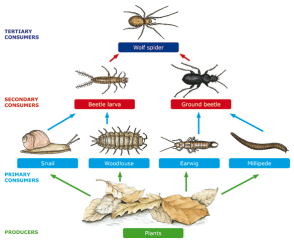

Figure 15: BLIP-2 answered 6/7 questions correctly with image descriptions and 3/7 without.

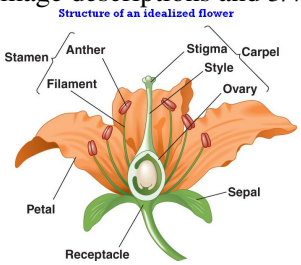

Figure 16: BLIP-2 answered 5/7 questions correctly with image descriptions and 2/7 without.

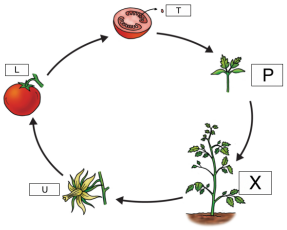

Figure 17: BLIP-2 answered 4/5 questions incorrectly with image descriptions and 5/5 without.

labeled images by generating image descriptions on the entire dataset, but due to time constraints we were unable to do that.

4.4.3 Qualitative Analysis of GPT-4V predictions on Diagram MCQs

GPT-4V's performance seems to be less random when it comes to image complexity. Our analysis indicated that it seems to struggle a bit with highly complex images that don't have image labels defined. For images that are slightly similar with a small number of labels, GPT-4V is able to do comparatively well as can be seen in Figure 19.

Figure 19, while having multiple leaves, is understood better by GPT-4V compared to the highly complex image in Figure 18 where GPT-4V has a 100% failure rate. We also think image clarity is important here since the image in Figure 19 is less cluttered with label values compared to Figure 18 where the image has many labels that are clustered together. There are around 210 questions

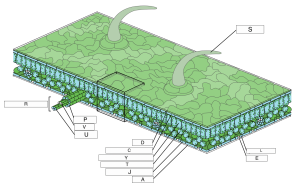

Figure 18: 4/4 questions answered incorrectly by GPT-4V

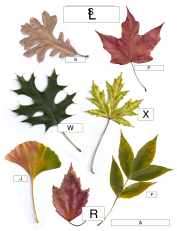

Figure 19: 5/5 questions answered correctly by GPT-4V

about images without defined labels that GPT-4V guessed the wrong answer for, and around 244 images without defined labels that GPT-4V guessed the right answer for. While the performance of GPT-4V suffers when label values are not provided, the reason for GPT-4V's high accuracy is because of the presence of defined labels in a large subset of images.

GPT-4V definitely has a slightly random performance on complex images as Figure 21 and 22 are as detailed as Figure 20, yet GPT-4V has a very low success rate on Figure 20 and an extremely high success rate on Figures 21 and 22 when answering very similar questions across both complex images. But, mostly, GPT-4V does very well on both simple and complex images by guessing answers correctly for 2251 out of 2831 images with defined labels. The degree of randomness in performance doesn't seem to be as high as that noticed for LLaVA-7B and BLIP-2 which shows that GPT-4V is able to generalize far better on Diagram MCQs from the educational domain.

5 Conclusion and Discussion

5.1 Replicability

Our results are quite easy to replicate and the only cumbersome portion is setting up the BLIP-2 and LLaVA packages to run locally on a high-end computer. Both BLIP-2 and LLaVA's GitHub pages contain many helpful example scripts and tutorials for running their model on a VQA task. It's quite simple to query GPT-4 as OpenAI's

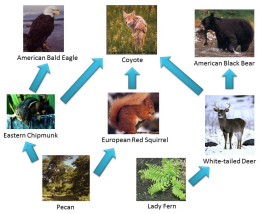

Figure 20: 6/7 questions answered incorrectly by GPT-4

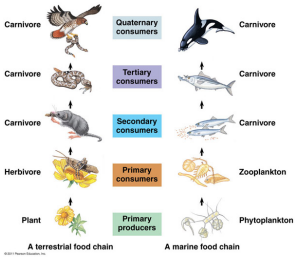

Figure 21: 7/8 questions answered correctly by GPT-4

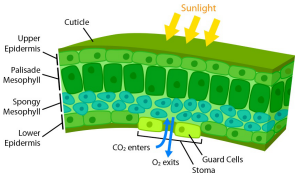

Figure 22: 7/9 questions answered correctly by GPT-4

website provides many helpful code snippets for making API requests with our custom zero-shot prompts.

5.2 Datasets

Our dataset was obtained from the AllenAI website and the only change we performed is parsing the json data obtained from the webiste to CSV files that could easily be processed using pandas for creating our zero-shot prompts. We did not create any extra annotations for the images in our dataset since they had already been annotated in 2017.

5.3 Ethics

We don't foresee any ethical issues associated with our case-study of modern VLM performance on the TQA dataset. All the data from the TQA dataset has been collected from middle-school curricula and doesn't contain any sensitive information about the schools or students/faculty that attend them. This is a regular case-study of a VQA task that we found interesting, and hence do not think there are any ethical issues with the way we

have carried out our analyses.

5.4 Limitation and broader impact

5.4.1 Few-shot, CoT and fine-tuning

One significant limitation this project faced is that we didn't have enough time to evaluate few-shot and CoT strategies, especially with BLIP-2 and LLaVA-7B. Hence, we were unable to verify Hypothesis III defined in subsection 2.1. However, since BLIP-2 and LLaVA-7B perform very randomly on our images regardless of their complexity, we believe that these strategies would be quite difficult to implement and would require careful thought as well as actual human-experts to collaborate and design meaningful prompts with. There also doesn't seem to be a clear pattern on what kind of questions both models fail on consistently, so it is definitely a challenging task to create few-shot prompts for these two models. Since these VLMs are so large, we were unable to fine-tune our models on the training set too - an approach that could turn out to be very useful for BLIP-2 and LLaVA-7B. GPT-4, particularly, has $\tilde{1}.76$ trillion parameters, which is a lot larger than the 2.7B BLIP-2 and LLaVA-7B parameter models, and thus we feel that GPT-4 and GPT-4V would definitely benefit most from few-shot or chain-of-thought reasoning.

5.4.2 GPT-4V generated image descriptions

In subsection 2.4, we mentioned how GPT-4V image descriptions were generated for a subset of images and were used in zero-shot prompts for BLIP-2 and LLaVA. However, most language models are prone to hallucinations and we didn't have any way of verifying if the image descriptions GPT-4 generated were actually thorough and accurate. We assumed GPT-4 generated accurate descriptions since it had ¿80% accuracy on questions associated with those images. The descriptions did help BLIP-2, but LLaVA only showed marginal improvements on the subset of images. Ideally, it would be great if we had human-expert generated image descriptions in the dataset, so that we could judge the performance of our augmented zero-shot prompts for all 3000+ images instead of just 403 images. We did generate captions using BLIP-2 but the generated captions were too short and were not very helpful.

5.4.3 Extending our methodology to other datasets

Furthermore, another limitation the project faced was that the models weren't tested with other datasets. The models were exclusively tested on the Textbook Question Answering Dataset. Given more time, we would have liked to judge our zero-shot performance on the ScienceQA dataset for GPT-4V using our custom prompts and compare it to Wu et al.'s results. The ScienceQA dataset contains similarly themed multi-modal questions as the TQA dataset that are derived from elementary and high-school curricula. Since the TQA dataset is mostly assembled from middle-school curricula, testing the zero-shot performance of modern VLMs like BLIP-2, LLaVA-7B and GPT-4V using our custom prompts would make for an interesting case study, specifically because Wu et al. were able to achieve a 85% zero-shot accuracy on the ScienceQA dataset (Wu et al., 2023) and we were only able to achieve a 76.7% zero-shot accuracy on the TQA dataset using GPT-4V.

6 References

- Kembhavi, Aniruddha, et al. "Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension." Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. 2017.
- Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv preprint arXiv:2301.12597 (2023).
- Liu, Haotian, et al. "Visual instruction tuning." arXiv preprint arXiv:2304.08485 (2023).
- Wu, Yang, et al. "An Early Evaluation of GPT-4V (ision)." arXiv preprint arXiv:2310.16534 (2023).
- Li, Dongxu, et al. "LAVIS: A one-stop library for language-vision intelligence." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). 2023.
- Liu, Haotian, et al. "Improved baselines with visual instruction tuning." arXiv preprint arXiv:2310.03744 (2023).