XAI: Prompting for Multimodal QA and Evaluation of Rationale Generation

Jatin Karthik Tripathy
University of Potsdam
tripathy@uni-potsdam.de

Supervised by: Dr. Sherzod Hakimov

Abstract

With the introduction of human-like conversational AIs in recent months, prompting has become the new paradigm for engaging with Large Language Models (LLMs). A few examples of input and output pairs within the model's context window appear enough for LLMs to infer how to perform a new language task on the fly without prior training. These models, for example, may perform arithmetic or answer trivia questions more accurately with a few context-relevant instances of important queries and proper solutions. Even if what may be concluded or learned from these inquiries is not always obvious. In this paper, we offer a new prompting approach for question answering and reasoning generating in Visual-Language data. We demonstrate a modular system for collecting the textual image context in three ways: visual tags, image captioning, and, lastly, combining both. We can observe the consequences of adjusting the number of examples presented to the LLMs using a prompt template. We further show, using a simple human annotation job, that applying typical NLP metrics to assess the produced rationales produces far poorer outcomes than is really the case.

1 Introduction

In recent months, there has been an advent of human-like conversational AIs, and with that, prompting has become the new paradigm for interacting with Large Language Models (LLMs). A few instances of input and output pairings within the model's context window seem to be sufficient for LLMs to infer how to carry out a new language task few-shot (Brown et al., 2020), without any prior training. With a few context-relevant examples of pertinent questions and appropriate responses, these models, for instance, may execute arithmetic or answer trivia questions more accurately. Even if it's not always evident what may be deduced or learnt from these questions (Min et al., 2022; Webson and Pavlick, 2021).

This lack of clarity can somewhat be understood and further explored by going a step further than just prompting to answer questions and also prompting for the explanation or the rationale behind the answer that the LLM generates. If we were to take a similar stance with LLMs as we do with humans, the act of justifying an answer should allow us to understand two main things: whether the task itself has been understood properly and whether there is any sense of coherency in the answer itself. As a result, explanations can help to explain the intended job by demonstrating the concepts that connect questions to answers.

Furthermore, while the overall performance of conversational AIs has only seen an increase, these systems are end-to-end text-based systems that directly map the input to the output. While this may be seen as a feature or a caveat depending on what the goal is, it is evident that tackling multimodal tasks require some workarounds. These workarounds usually tend to involve a preliminary step in translating the multi-modal data into text so that the end-to-end nature of LLMs can be preserved, as seen in the work done by Cheng et al. (2022).

That said, some works take differing approaches to handling multi-modal data. For example, we could also be capable of extending the LLMs themselves by joining them with models that are capable of handling the other modality (Park et al., 2018; Wu and Mooney, 2019; Sammani et al., 2022; Palaskar et al., 2022). The primary issue with such unified models, however, is the fact that they need to be retrained or finetuned to some extent to ensure that the different modalities are sufficiently linked and can be used for QA and rationale generation.

Thus, the aim of this work is two-fold: first, to investigate whether an image can be accurately described via text to allow LLMs to answer questions based on the image in a few-shot manner and second, to investigate if the LLMs are then capable of

generating rationale that can support the previously generated answer.

This aim is interesting due to multiple reasons. By translating the different modalities into text, we can explore many more different methods of approaching this problem without being limited by retraining a model. This allows us to simplify the overall problem as the complexity of the task goes down to just converting an image to text rather than having a single model be responsible for QA and rationale generation. Furthermore, practically this allows us to build up a pipeline that allows for different state-of-the-art models to be swapped and tested.

2 Related Works

As mentioned in the previous section, there are two main approaches when it comes to dealing with visual-language question answering and rationale generation tasks. One is an end-to-end unified approach, while the other tackles the image and language data separately. In recent years, several works have been proposed in both methods, having their own advantages and disadvantages.

One of the more recent end-to-end unified approaches, NLX-GPT (Sammani et al., 2022) unites response prediction and explanation creation by training a simplified version of the GPT-2 Transformer decoder. The decoder predicts a series of questions, answers, and explanations as the prediction target. The encoded picture is delivered to the GPT-2 decoder's cross-attention layer as the value and key vectors. At inference time, the decoder accepts the query as input and produces the response and explanation autoregressively. NLX-GPT, on the other hand, has not been trained on a wide range of tasks, necessitating the use of a task-specific module.

Palaskar et al. (2022) introduces VL-T5 and VL-BART, which expand the Transformer-based unimodal language models BART-Base (Lewis et al., 2020) and T5-Base (Raffel et al., 2020) with visual inputs. The two models, like NLX-GPT, combine response prediction with explanation creation. However, they have the same problem in that they have not been finetuned on various tasks, limiting task accuracy and explanation plausibility ratings.

On the other hand, Kayser et al. (2021) modified the RationaleVT Transformer model given by Marasovic et al. (2020), who merged several vision-language models with a fine-tuned GPT-2 language model for rationale generating. Instead of the vision-language models presented by Marasovic et al. (2020), e-UG predicts a task response label using the UNITER vision-language Transformer as an answer prediction module. The representations used to forecast the label are then sent into the GPT-2 rationale module, together with additional response and question text input, to construct the justification. Both e-UG and RationaleVT Transformer remove the NLE generating module from the answer prediction module, resulting in larger models with a larger number of parameters and poor performance.

RExC (Majumder et al., 2022) also takes on a complicated modular approach: it first utilizes a vision-language model to extract the bounding boxes of the important objects in the picture and then uses their characteristics to build knowledge snippets using a knowledge module. Finally, the rationale generation module is given relevant knowledge snippets, and the resulting concealed vector is passed to the prediction module. By definition, this highly modular approach is dependent on several external models. This, once again, quickly necessitates a larger number of parameters and separate maintenance of the individual modules when applying the model to new data.

More recently, the more common approach that is being followed is to create multitask multimodal models that have been trained on various datasets. By introducing visual information into the LLMs, instruction-tuned LLMs have been applied to image-to-text creation. BLIP-2 (Li et al., 2023) use frozen FlanT5 models as input to the LLMs and trains a Q-Former to extract visual characteristics. MiniGPT4 (Zhu et al., 2023) has the same pre-trained visual encoder and Q-Former as BLIP-2 but employs Vicuna as the LLM and trains using lengthier picture captions than BLIP-2. LLaVA (Liu et al., 2023) immediately projects the output of a visual encoder as input to a LLaMA/Vinuca LLM and finetunes the LLM using GPT-4 (OpenAI, 2023) vision-language conversational data. mPLUG-owl (Ye et al., 2023) uses low-rank adaption to fine-tune a LLaMA (Touvron et al., 2023) model utilizing both text instruction data and LLaVA vision-language instruction data.

However, the main drawback of these newer end-to-end multitask multimodal models is that they need to be finetuned or trained on a large amount of data before they can be used. Thus in this work,

we shall endeavour to create a more modular approach like RExC, which will not only allow both the vision and language models to be decoupled but also permit the use of few-shot inference using the existing pretraining.

3 Methodology

As mentioned in Section 1, the primary goal of this work is twofold. The first challenge that we tackle is to be able to accurately describe an image using only text, with this being done in three different ways and will be discussed in Section 3.1. Once an image is transcribed to a textual format, they then do the question-answering stage of the task. Following this is the second challenge that this work looks at, which is the attempt to generate a suitable rationale behind the generated answer. Since this work takes advantage of the more recent text-to-text LLMs, we also have the need to wrap up the whole process and frame it as a prompt for the LLMs to be able to understand and produce sensible outputs. The prompt template and the reasoning behind the structuring of the template will be discussed in Section 3.2 This whole process can be aptly summarized as seen in Figure 1.

3.1 Visual Context Generation

Obtaining the context of an image as accurately and detailed as possible is integral to this work, as without having proper context, it is impossible even for humans to look at just the textual form of the image context and answer the question. Thus, we employ three different strategies to experiment with the best way possible to get the image context: Visual Tagging, Captioning, and finally, Combining both the Visual Tags and the Captions, as seen in Figure 2.

3.1.1 Visual Tags

In this approach, we try to keep things as modularized as possible by employing different image models to focus on specific aspects of the image itself. This method allows us to try and work around the fact that some context is harder to grab than others, and by splitting the workload, the generated textual image context can be richer than if we were to use a captioning model. The main caveat here is that this method is only as good as the number of individual image models that we use - more models do produce a better textual image context. Still, it comes at the cost of needing to run the image through several models. In this work, we use five different image models to obtain the context.

`Object Detection - DETR ResNet-101:` The DETR model (Carion et al., 2020) is a convolutional backbone encoder-decoder transformer. To detect objects, two heads are added to the decoder outputs: a linear layer for the class labels and an MLP (multi-layer perceptron) for the bounding boxes. To detect items in a picture, the model uses object queries. Each object query searches the picture for a specific object. The number of object queries for COCO is set to 100.

The model is trained using a "bipartite matching loss": the predicted classes + bounding boxes of each of the N = 100 object queries are compared to the ground truth annotations, which have been padded up to the same length N. To establish an optimum one-to-one mapping between each of the N queries and each of the N annotations, the Hungarian matching method is utilized. The model's parameters are then optimized using conventional cross-entropy (for the classes) and a linear combination of the L1 and generalized IoU loss.

`Indoor Scene - MIT ViT-base:` The Vision Transformer (ViT) (Dosovitskiy et al., 2021) is a supervised transformer encoder model (BERT-like) that was pre-trained on a huge collection of pictures, specifically ImageNet-21k, at a resolution of 224x224 pixels. The model was then fine-tuned using ImageNet (also known as ILSVRC2012), a dataset of 1 million pictures and 1,000 classes at 224x224 resolution. The model is fed images as a series of fixed-size patches (resolution 16x16) that are linearly embedded. A [CLS] token is also added to the beginning of a sequence to be used for classification tasks. Before sending the sequence to the Transformer encoder layers, absolute position embeddings are also included.

We use a ViT model pre-trained on the MIT indoor scene dataset. The database contains 67 Indoor categories and a total of 15620 images. The number of images varies across categories, but there are at least 100 images per category.

`Outdoor Scene - Places365 ResNet50:` ResNet 50 (He et al., 2016) is a convolutional neural network that has made residual learning and skip connections more accessible. This allows for considerably deeper models to be trained. ResNet-50 is built on the original ResNet-34 design but with one significant variation. The bottleneck design is used for the building block in the 50-layer

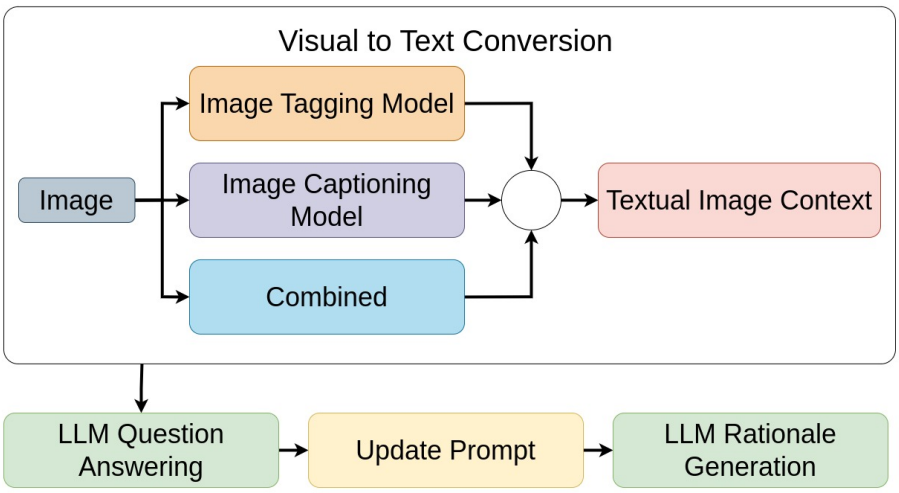


Figure 1: Pipeline

ResNet. A bottleneck residual block employs 11 convolutions to limit the number of parameters and matrix multiplications. This allows for significantly quicker layer training. It employs a three-layer stack rather than two levels.

For this work, we use a ResNet-50 model pre-trained on the Places365 dataset. Places-365 features over 10 million photos divided into 400+ scene types. The dataset includes 5000 to 30,000 training photos per class, which corresponds to the real-world frequency of occurrence. The Places dataset, which employs convolutional neural networks (CNN), enables the learning of deep scene characteristics for diverse scene identification applications.

`Face Detection - MTCNN:` We use a pre-trained Multi-task Cascaded Convolutional Networks (MTCNN) (Zhang et al., 2016) model designed to solve both face identification and face alignment problems. The method comprises three levels of convolutional networks capable of recognizing faces and landmark locations such as the eyes, nose, and mouth. The MTCNN must produce three results: face/non-face classification, bounding box regression, and facial landmark localisation.

`Emotion Detection - FER ViT-base:` We again use a ViT-base model, this time pre-trained on the Facial Expression Recognition 2013 (FER-2013) dataset. The data comprises of grayscale pictures of faces 48x48 pixels in size. The faces have been automatically registered such that the face is more or less in the centre of each image and occupies roughly the same amount of area. The aim is to identify each face into one of seven groups depending on the emotion expressed in the facial expression. The training set has 28,709 cases, whereas the public test set contains 3,589 examples. We also do a step of preprocessing by cropping out only the faces returned from the MTCNN model rather than sending the whole image for emotion detection as cropping produced far more reliable results.

3.1.2 Captioning

Unlike the previous approach, by using a more general-purpose Visual-Language model, we avoid the issue of having the need to manually select which visual tags to look out for in the image. This also ensures that the textual image context that is generated is more readable and forms a proper sentence rather than the context generated by the previous method. For this work, we make use of an implementation of the InstructBLIP (Dai et al., 2023).

The InstructBLIP model is an instruction-aware visual feature extraction method that allows for flexible and informative feature extraction based on the instructions provided. The command is supplied not only to the frozen LLM as a prerequisite for producing text, but it is also given to the Q-Former as a condition for extracting visual characteristics from the frozen image encoder.

3.1.3 Combined

The primary advantage that this method has over the other two is the fact that we gain the advantage of specifically targeting specific visual cues while having a more general approach to the captions themselves. To keep things fairly simple, the

Gold Caption: *a plate with pizza, lettuce and ham sit on a plate white we see hands holding silverware and a bottle of wine with glasses*

Visual Tags: *This is an image with person, wine glass, bottle, pizza, spoon, fork, dining table*

Caption: *a white plate topped with a pizza covered in toppings*

Combined: *This is an image with person, wine glass, bottle, pizza, spoon, fork, dining table, a white plate topped with a pizza covered in toppings*

Figure 2: Example of different Visual Context Generation methods

combination method was kept to just appending the caption to the end of the context generated while using the Visual Tagging method. While this method does often cause a lot of overlap in the information and sometimes leads to the textual image context being long, the LLM prompting tends to be more robust, as seen in Section 6.

3.2 Prompt Template

For the actual prompting of the LLMs themselves, we use a prompt template as shown in Figure 3. Brown et al. (2020)'s work showed that LLMs are quite capable few-shot learners that, in most cases, do not require any further fine-tuning required. They also showed that with just prompting, it is possible to get use natural language description of the task to make the model understand and provide reliable outputs.

At inference time, the model is given a few examples of the task as conditioning, but no weights are updated. An example usually has a context and an intended outcome. The key advantage of a few-shot is that it significantly reduces the demand for task-specific data.

While the main downside is that the outcomes of this technique have so far been significantly inferior to those of state-of-the-art fine-tuned models. A limited quantity of task-specific data is also necessary. As the name implies, few-shot learning for language models is linked to few-shot learning used in other situations in NLP.

In this work, we make use of this design paradigm and experiment by including N in-context examples of the task as well as trying a zero-shot approach. The prompt header is a one or two-sentence description of the task itself. We noticed that without the inclusion of the prompt header, the models become quite unstable even if we were to provide N in context samples for the model to "learn" from.

The prompting of the LLMs itself is a very linear process as described in Section 3. We iteratively append in-context samples to the prompt header, with the in-context samples having all the necessary pieces of information: the textual image context, the question and the answer.

For the actual answer prompting, the only difference between this and the in-context sample is that the answer is left blank for the model to fill out. Similarly, for the rationale prompting, we append the generated answer back to the prompt and, this time, leave the rationale part black for the model to fill.

We did notice that the inclusion of the rationales in the in-context samples does not seem to affect either the generated answer or the generated rationale much. As such, in the prompt template that this work uses, we do not include the rationale to ensure that the prompt does not get too long. In the case of text only datasets, the only difference in the prompt itself is that the `Image Context` is not included.

4 Datasets

In this work we make use of four different datasets. Two of which, VQA-X (Park et al., 2018) and A-OKVQA (Schwenk et al., 2022), are Visual-Language multimodal datasets and are the main topic of interest in this work. We also test on two text-only datasets, SENMAKING (Wang et al., 2019) and e-SNLI (Camburu et al., 2018), to act as a baseline to show how the LLM models perform while prompting. They let us compare the results obtained to check the feasibility of prompting using the textual image context.

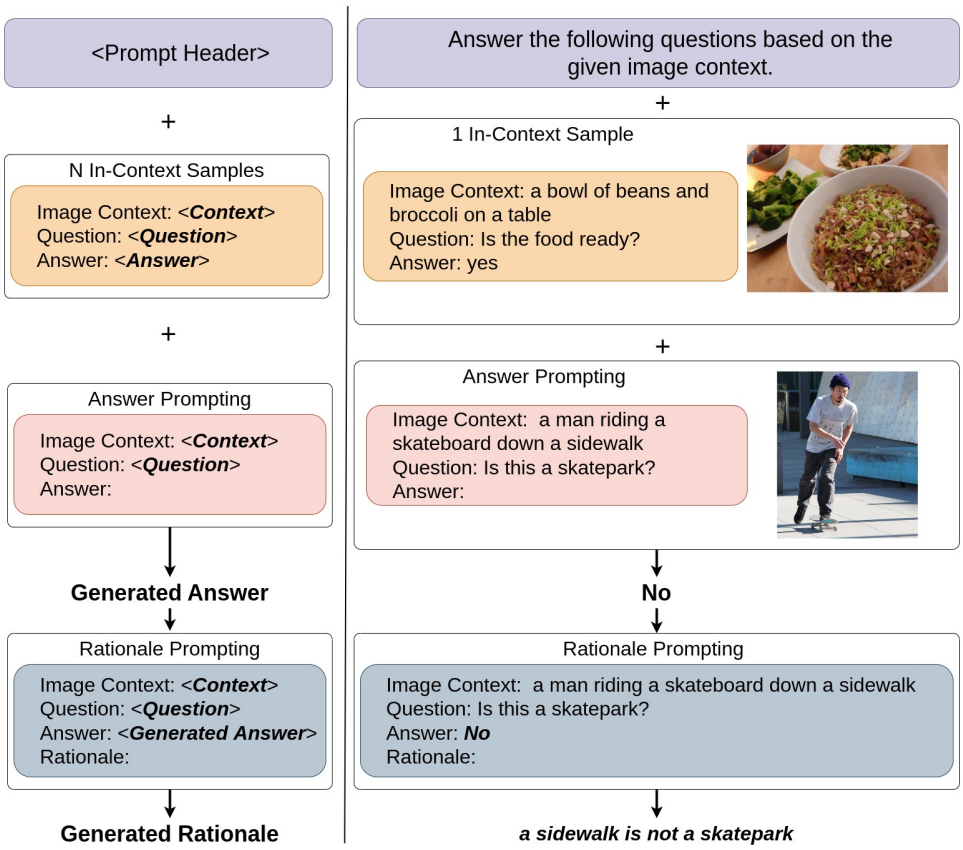


Figure 3: Prompt Template and Example. Image context in the example is obtained using a captioning model.

4.1 VQA-X

VQA is a well-studied multimodal task that includes visual and textual comprehension as well as common sense. The newly gathered VQA v2 dataset [16] contains question-and-answer pairings that are complimentary. Complementary VQA pairs pose the same question to two semantically comparable pictures with opposing responses. Because the two photos are semantically similar, VQA models must use fine-grained reasoning to appropriately answer the question. This is an intriguing and valuable setting not just for assessing overall VQA performance but also for examining explanations. We may more quickly identify whether our explanations focus on the crucial variables for making a decision by comparing explanations from complimentary pairings. The VQA-X dataset specifically focuses on visual question answering.

Question What is the person doing?

Answer Snowboarding

Rationale Because they are on a snowboard in snowboarding outfit.

Figure 4: **VQA-X Prompt Header:** Answer the following questions based on the image context

4.2 A-OKVQA

A-OKVQA is a crowd-sourced dataset comprised of around 25K multiple-choice questions that need a broad basis of commonsense and global knowledge to answer. Unlike the previous knowledge-based VQA datasets, the questions often cannot be answered by merely querying a knowledge base but rather involve some commonsense reasoning about the situation portrayed in the image.

They created A-OKVQA using photos from the 2017 partitioning of the COCO dataset since it includes numerous images congested with various items and is an established dataset with several re-

lated models already in existence. The questions in A-OKVQA were prepared and improved by 437 crowd-workers on the Amazon Mechanical Turk platform during numerous rounds of annotation. Workers completed a qualification task to demonstrate their ability to write questions that met our criteria, which include: (1) looking at the image to answer, (2) some commonsense or specialized knowledge, (3) some thinking beyond simply recognizing an object, and (4) not being too similar to previous questions.

After questions and multiple-choice answer options were collected and validated, They initiated a task to collect rationales. Workers were given question and response possibilities, and they were asked to explain in one to two short phrases why a certain answer was accurate, including any relevant information or knowledge about the world that was not included in the visuals. To ensure high-quality production, workers were provided examples and went through a qualifying procedure with three rationales gathered for each query.

Question How many people will dine at this table?

Answer one

Rationale There is only one cup of water and main dish at this table.

Figure 5: **A-OKVQA Prompt Header:** Answer the following questions based on the image context

4.3 SEN-MAKING

The dataset is divided into two subtasks: the first is to determine which of two natural language statements with identical wordings makes sense and which does not; the second is to determine the fundamental reason why a particular statement does not make sense. It has 2021 examples for each subtask that 7 annotators have carefully tagged. On the benchmark, human performance is 99.1% for the Sen-Making job and 97.3% for the Explanation test.

Statement 1: He put a turkey into the fridge

Statement 2: He put a elephant into the fridge

Choice A: an elephant cannot eat a fridge

Choice B: elephants are usually gray while fridges are usually white

Choice C: an elephant is much bigger than a fridge

Figure 6: **SEN-MAKING Prompt Header:** Which statement of the two is against common sense?

Annotators were instructed to follow many rules when composing samples. To begin, avoid difficult information and focus on everyday common sense, and keep the questions as simple as feasible. Every literate individual may provide the correct answers. Second, under the against-commonsense claims, the puzzling reasons should include more key terms, such as things and actions. For example, the confusing reasons "he put an elephant into the fridge" should include both "elephant" and "fridge." Third, we want the confused reasons tied to the assertions and proper reasons while remaining inside the problem context. They also regulate the length of each phrase, such that the inaccurate statement is approximately as long as the correct statement, and the proper explanation is neither too long nor too short of the three reasons.

4.4 e-SNLI

Premise: An adult dressed in black holds a stick.

Hypothesis: An adult is walking away, empty-handed.

Label: contradiction

Explanation: Holds a stick implies using hands so it is not empty-handed.

Figure 7: **e-SNLI Prompt Header:** What is the relation between the two sentences. Possible answers are: contradiction, entailment, and neutral.

They add a layer of human-annotated natural language explanations of entailment connections to the Stanford Natural Language Inference dataset. They then build models that include these explanations in the training process and output them during

testing.

The fundamental question they want our dataset to answer is, "Why is a pair of phrases in an entailment, neutrality, or contradiction relationship?". They asked the annotators to pay attention to the non-obvious factors that cause the provided relationship rather than the portions of the premise that are reproduced exactly in the hypothesis. They required arguments for those aspects of the hypothesis that did not exist in the premise for entailment. While they advocated mentioning all of the aspects that contribute to the relationship for neutral and conflicting pairings, they consider an explanation correct if at least one element is given. Lastly, they requested the annotators to produce self-contained explanations rather than words that made sense only after reading the premise and hypothesis.

5 LLMs

5.1 Dolly v2-12b

Dolly v2 works by adopting data from Alpaca to enhance an existing GPT-J-6b model from EleutherAI in order to elicit "instruction following" features like brainstorming and text production that were not included in the original model. Dolly's underlying model contains only 6 billion parameters, compared to 175 billion in GPT-3. This shows that targeted corpora of instruction-following training data, rather than bigger or better-tuned base models, may account for a major portion of the qualitative advances in cutting-edge models like ChatGPT.

5.2 FLAN T5-xxl

The FLAN-T5-xxl (Raffel et al., 2020) model is a pre-trained encoder-decoder model with 11 billion parameters that was trained on a multi-task combination of unsupervised and supervised tasks, with each task transformed into a text-to-text format. FLAN-T5 was fed a huge corpus of text data during the training phase and taught to predict missing words in an input text using a fill-in-the-blank type goal. This procedure is continued until the model learns to create text that is comparable to the input data. Packing (Raffel et al., 2020) merges several training samples into a single sequence, with inputs and targets separated by an end-of-sequence token. To prevent tokens from attending to others over the packed example border, masking is used.

5.3 T0pp

T0pp (Sanh et al., 2022) is an 11 billion parameter model that demonstrates zero-shot task generalization on English natural language prompts, exceeding GPT-3 on a variety of tasks while being 16x smaller. It is an encoder-decoder model that has been trained on a diverse variety of natural language tasks. They turn a large number of English-supervised datasets into prompts, each with different templates and formulations that vary. These prompted datasets enable measuring a model's ability to accomplish previously unknown tasks described in normal language. To get T0pp, they fine-tune a pre-trained language model on this multitask mixture that covers a wide range of NLP activities.

6 Results

The results obtained from the experiment this work undertook is quite interesting, given in Table 1. Across all four datasets, we can see that the performance of Dolly v2 and T0pp is quite bad, and even more interestingly, both the LLMs tend to get worse as more in-context samples are added. On the other hand, FLAN T5-xxl showcases expected behaviour, with the LLM performing ever so slightly better when adding more in-context samples. This could be because of the pretraining method FLAN T5 uses, a more text-to-text formatting of all task when compared to the pretraining methods of the other two LLMs.

Looking more closely, we can also see a trend when comparing the different methods of obtaining the textual image context. Using only the visual tags makes the LLMs perform the worst across the board, with the exception of Dolly v2, see Appendix A. When comparing the different methods of textual image context in FLAN T5-xxl, we also see that the combined method allows the LLM to ground itself more and obtain better context. This would then imply that while using only visual tags does not seem to be enough, they do provide extra information that the caption alone can not obtain.

When looking at the rationale generation aspect of this work, we see that both models fail quite abysmally when looking at the metrics. With only FLAN T5-xxl managing to get an understandable text on the A-OKVQA dataset and still failing in the other Visual-Language dataset. As expected from the Accuracy scores, FLAN T5-xxl is able to achieve quite high scores in metrics in both of the text only datasets.

Table 1: Best Accuracies

Dataset	Models	Visual Context	N In-Context	Accuracy	BLEU-4	ROUGE	METOER
	Dolly-v2-12b	Visual Tags	1	0.085	0.007	0.122	0.211
VQA-X	FLAN T5-xxl	Combined	4	0.668	0.077	0.316	0.326
	T0pp	Caption	1	0.091	0.063	0.284	0.246
	Dolly-v2-12b	Visual Tags	0	0.064	0.028	0.211	0.295
A-OKVQA	FLAN T5-xxl	Both	3	0.646	0.855	0.287	0.289
	T0pp	Caption	0	0.577	0.089	0.300	0.287
	Dolly-v2-12b	-	-	0	-	-	-
SEN-MAKING	FLAN T5-xxl	-	4	0.971			
	T0pp	-	-	0	-	-	-
	Dolly-v2-12b	-	3	0.042	0.117	0.097	0.188
e-SNLI	FLAN T5-xxl	-	2	0.861	0.018	0.092	0.110
	T0pp	-	0	0.282	0.088	0.269	0.249

Image Context: a woman sitting in front of a christmas tree

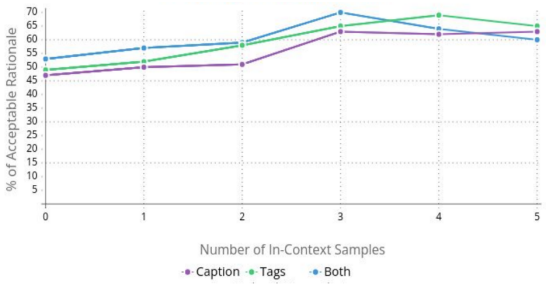
Question: What holiday is being celebrated?

Generated Answer: Christmas

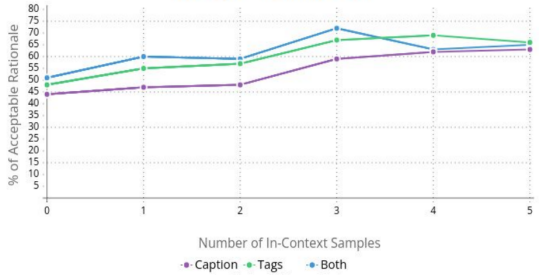
Gold Rationale: She has presents in her hand

Generated Rationale: Christmas is the holiday that is celebrated with a Christmas tree

Figure 8: Example of a prompt that is impossible to answer



A-OKVQA



VQA-X

Figure 9: Caption

However, this discrepancy in the Visual-Language dataset is more of a fallacy with any reasoning task. Like we can see in Figure 8, while the generated rationale is in no way similar to the gold standard rationale, we can not say that the generated rationale is wrong. But due to the method by which most NLP metrics work, these alternate rationales generated by the LLMs are deemed to be wrong, and the overall score goes down.

To try and combat this, we hand annotate 100 samples of the rationale generated by FLAN T5-xxl in both of the Visual-Language datasets and check if the rationales are truly wrong or whether they can be accepted. To our pleasant surprise, we found that most of the generated rationales are quite acceptable, around 70%, if very different from the gold standard rationales present in the datasets, as shown in Figure 9.

A common point of failure for all the tested LLMs is when the textual image context does not grab the information needed to answer the question, Figure 10. In these cases, the LLMs tend to make completely random guesses based on the question which then leads to the generated rationale also being quite off the mark when compared to the gold standard rationale. In the cases where the generated rationale is not on point, we also see a phenomenon where the model tends to paraphrase the caption to

Image Context: This is an image of 3 person, water, pier, a group of people standing on top of a pier next to a body of water

Question: Is it daytime?

Generated Answer: No

Generated Rationale: the people are standing on top of a pier

Figure 10: Example of a prompt that is impossible to answer

form the rationale.

7 Conclusion

In this work, we propose a new prompting method for question answering and rationale generation for Visual-Language data. We show a modular method of obtaining the textual image context in three different ways: Visual Tags, Image captioning, and finally, combining both. By using a prompt template, we are able to incorporate in-context samples, thus letting us see the effects of changing the number of examples provided to the LLMs. We also show using a small human annotation task that using traditional NLP metrics to evaluate the generated rationales tend to show much worse results that what is actually the case. Finally, this work shows that both the question answering and the rationale generation using FLAN T5-xxl accomplishes the original question of whether it is possible for conversational LLMs to comprehend Visual-Language data in a contextually textual form.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Binding language models in symbolic languages. *CoRR*, abs/2210.02875.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *CoRR*, abs/2304.08485.

Bodhisattwa Prasad Majumder, Oana Camburu, Thomas Lukasiewicz, and Julian J. McAuley. 2022. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 14786–14801. PMLR.

Ana Marasovic, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2810–2829. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Shruti Palaskar, Akshita Bhagia, Yonatan Bisk, Florian Metze, Alan W. Black, and Ana Marasovic. 2022. On advances in text generation from images beyond captioning: A case study in self-rationalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2644–2657. Association for Computational Linguistics.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8312–8322. IEEE.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? A pilot study for sense making and explanation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4020–4026. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *CoRR*, abs/2109.01247.

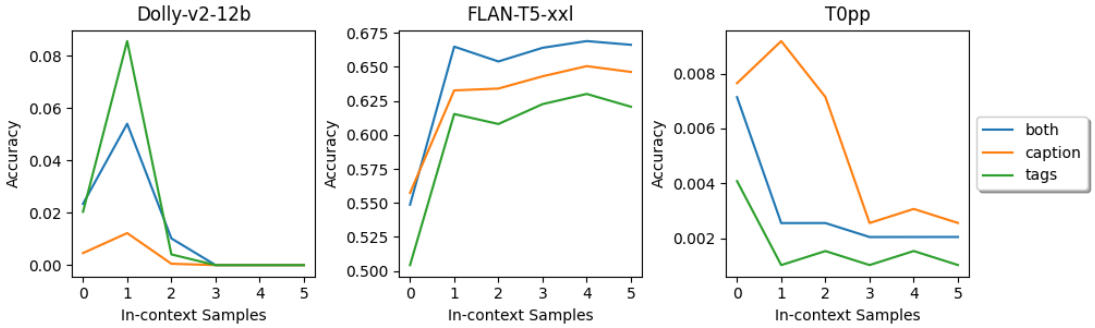
Jialin Wu and Raymond J. Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 103–112. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

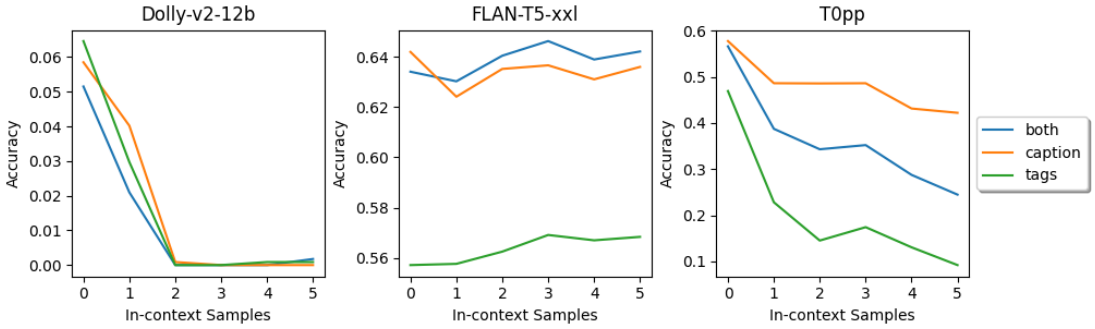
Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

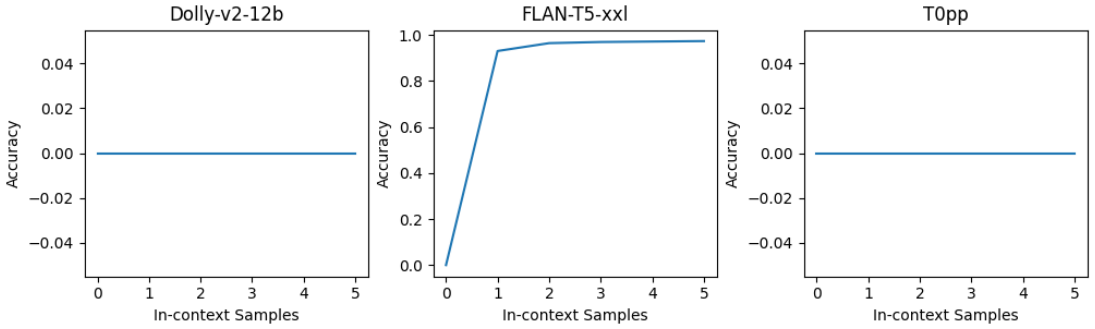
A Accuracies of Different LLMs across all Datasets



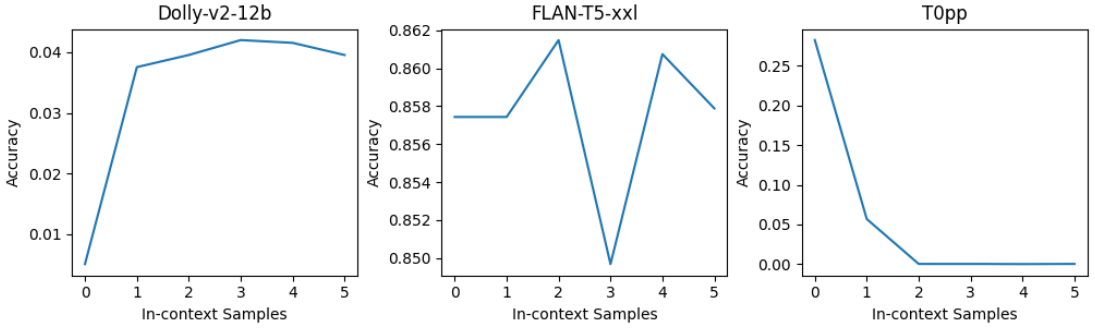
VQA-X



A-OKVQA



SEN-MAKING



e-SNLI