MAEV: Multimodal Automatic Evaluation Framework

Anonymous EMNLP submission

Abstract

We propose a novel framework (MAEV), designed to evaluate the performance of multimodal large language models (MM-LLMs) on complex, open-ended multimodal reasoning tasks. Our approach mitigates the limitations in the conventional classification-based MM-LLM evaluation methods, providing a comprehensive analysis of free-form MM-LLM responses by leveraging state-of-the-art LLMs. To achieve this, we introduce two carefully crafted evaluation datasets comprising $2K$ ground-truth long-form responses to open-ended visual queries and detailed image descriptions. Our experimental results demonstrate the effectiveness of MAEV, as it closely aligns with human evaluation outcomes and offers a much-needed solution to complement the time-consuming manual assessment process. This framework has the potential to accelerate the development of cutting-edge MM-LLMs.

1 Introduction

Multimodal Large Language Models (MM-LLMs) (Alayrac et al., 2022; Awadalla et al., 2023; Zhu et al., 2023; Liu et al., 2023; Moon et al., 2023) have gained increasing popularity in recent years, due to their ability to reason over image and text queries, especially since the introduction of GPT4V. However, automatic evaluation of MM-LLMs is typically based on a limited set of VQA benchmark datasets, which may not accurately reflect the model's true performance. Recent studies have shown that there is a significant discrepancy between classification-based automatic evaluation and human evaluation on real-world open-ended queries (Moon et al., 2023) – *e.g.* models that perform similarly on short-form VQA tasks may fare much worse when evaluated by humans. This is because VQA benchmarks often consist of concise and simple answers, which do not capture the full range of a MM-LLM's capabilities.

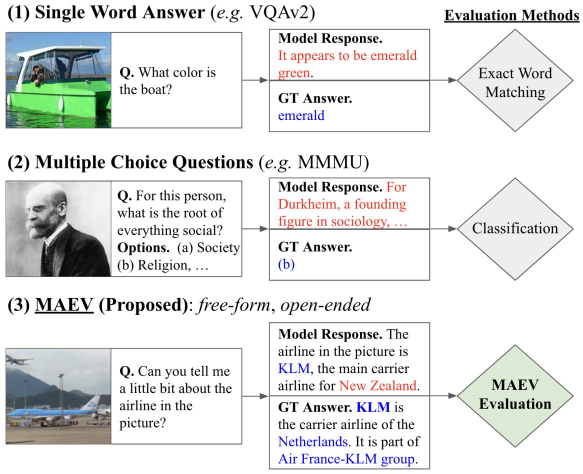

Figure 1: Comparison of different evaluation methods for MM-LLMs. (1) Single-word answer questions (*e.g.* VQAv2 (Antol et al., 2015)) and (2) multiple choice questions (*e.g.* MMMU (Yue et al., 2023)) typically concern evaluation of object- or attribute-level understanding, or are limited in the depth of questions by their deterministic nature. In addition, due to their output format, fluent MM-LLM responses often get penalized unnecessarily. (3) Our proposed MAEV offers a more comprehensive evaluation approach, enabling the assessment of free-form and open-ended queries, and providing fine-grained feedback on the detailed aspects of long-form model responses.

Manual evaluation is time-consuming and costly, which hinders the fast iteration and development of MM-LLMs. To address this issue, we propose a novel framework called Multimodal Automatic Evaluation (MAEV), which is a model-agnostic approach for evaluating MM-LLMs on challenging multimodal reasoning tasks (Figure 5). Specifically, we construct new gold standard datasets that contain dense annotation of visual information, such as ground-truth captions and assistant responses, to enable automatic judgment of model responses using *text-only LLMs as the evaluator*. Note that this dense annotation removes the dependency on a secondary multimodal model (with presumably similar visual understanding capabilities as the models being evaluated), which defeats the purpose. We

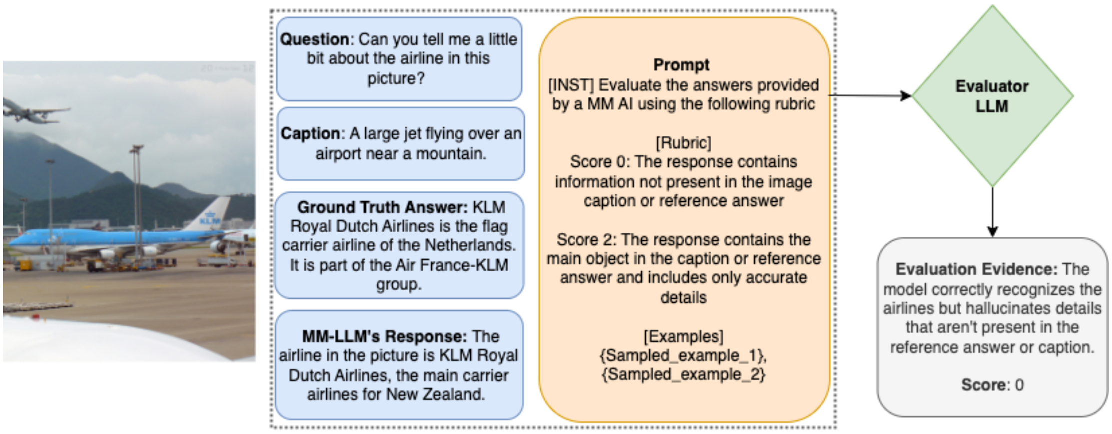

Figure 2: Illustration of the Evaluation flow of MAEV. The image caption serves as a proxy for the image and is fed into the evaluator LLM instead of the image itself. Appropriate examples are sampled for each trial run, producing evaluation evidences and appropriate scores.

then design a statistical framework to measure the performance of MM-LLMs reliably for each given task sample with varying few-shot prompts, and make a holistic evaluation of a given model.

We conduct case studies comparing the performance of two MM-LLMs under development, using both manual human evaluation and MAEV. Our results show that MAEV successfully tracks the human evaluation results on pointwise evaluation, demonstrating the feasibility of replacing or complementing manual evaluation with MAEV.

2 Related Work

LLM-Based Evaluation: Recent advancements in LLMs have opened new avenues for text evaluation, proposing alternatives to traditional human assessment methods. Studies such as Chiang and Lee (2023) have explored the potential of LLMs as substitutes for human evaluation in NLP tasks. They demonstrate a high degree of consistency between LLM evaluations and expert human ratings, especially in complex tasks like open-ended story generation and adversarial attacks. Additionally, Zheng et al. (2023) extended the use of LLMs to the evaluation of conversational AI systems, introducing benchmarks that showcase their effectiveness in aligning with human preferences while addressing inherent biases and limitations.

Challenges and Advancements in LLM Evaluation: While the potential of LLMs as evaluators is evident, significant challenges and biases remain. Wang et al. (2023) brought to light the biases in LLM evaluations, particularly how response rankings can be influenced by the order of presentation, emphasizing the need for mitigation strategies and ethical considerations. In response, Kim et al. (2023) introduced Prometheus, a fine-tuned LLM showcasing its effectiveness as an evaluator with capabilities comparable to GPT-4 and a strong correlation with human evaluators, marking a significant advancement in LLM evaluation.

Our work extends the previous line of LLM-based automatic evaluation framework of text-only benchmarks to the multimodal settings, with a new set of multimodal evaluation datasets.

3 Methods

This study introduces a novel framework for evaluating MM-LLMs using (1) a text-based LLM as the evaluator (Sec.3.1) on a (2) newly constructed evaluation datasets with dense visual information fully annotated (Sec.3.2). The primary focus of this evaluation is to assess the accuracy of the MM-LLM's responses to queries that incorporate both textual and visual inputs.

3.1 LLM as Evaluator Model

The evaluation model is conducted using an Evaluator LLM that analyzes the outputs of MM-LLMs. This process takes in the following inputs:

Text Query: questions provided to the MM-LLM.

Image Caption: A textual representation of the image, serving as a surrogate for the actual image.

Ground Truth Answer: The correct response to the query, used as a benchmark for accuracy.

MM-LLM Response: The actual response generated by the MM-LLM to the query.

The primary criterion for evaluation is to measure the accuracy, specifically assessing if there are any hallucinations in the MM-LLM's response as compared to the ground truth answer, or hallucinations in recognizing the objects described in the image captions.

3.2 Eval Data w/ Dense Visual Annotation

A comprehensive and representative dataset is compiled for this evaluation, encompassing various categories such as plants, text, sightseeing, and landmarks. The dataset is prepared through the following steps:

Image Collection: Gathering a diverse set of images from the specified categories.

Caption Creation: Human annotators creating descriptive text captions for each image, effectively translating the visual content into text.

Question and Answer Formulation: Annotators developing questions that a human observer might naturally ask about each image, along with the appropriate ground truth answers.

3.3 Evaluation Procedure

The evaluation procedure for the MM-LLM was designed as follows:

Dataset Creation: The prepared dataset, including images, captions, questions, and ground truth answers, is employed for the evaluation.

MM-LLM Response Generation: The MM-LLM is presented with each image and the corresponding question to generate a response.

Evaluation Input: The Evaluator LLM receives the set of inputs - the question, image caption, ground truth answer, and MM-LLM response for each sample task.

Scoring Methodology: Each response from the MM-LLM is scored based on its accuracy relative to the ground truth. A score of 0 indicates the presence of hallucination and a score of 2 indicates a correct response.

Output: The output of the evaluation is first the evaluation evidence followed by the evaluation score. The evidence is vital to explain model behaviour. Additionally, by providing the evidence first we take advantage of the auto-regressive nature of LLMs to provide a rational score.

To improve the evaluator accuracy and robustness, we apply the following strategies

Few-Shot Sampling: In our evaluation framework, few-shot learning is employed to enhance the MAEV evaluator accuracy. Each prompt is supplemented with two examples, selected from a larger pool of ten, to guide the MM-LLM's response. This approach acknowledges the context length constraints in each evaluation. The choice of examples is pivotal for a balanced assessment. We select one example with a score of 0, showing complete hallucination. Another example is chosen with a score of 2, reflecting high accuracy.

Multi-Trial Aggregation: To account for the inherent variability in language model responses, we utilize a multi-trial aggregation method. Each trial samples a different pair of examples (scored 0 and 2) for prompting the MM-LLM. By aggregating outcomes from multiple trials, we ensure a more balanced and generalizable assessment, reducing biases that may result from a single example set.

4 Experiments

4.1 Dataset

We employ two distinct types of image sources:

COCO Dataset: We sample 800 images from the COCO dataset (Lin et al., 2014), and collect new free-form question and answer annotations. COCO is integral to our research for its diversity in scenes and objects, along with comprehensive annotations. It is particularly suitable for tasks involving object detection and segmentation, offering a plethora of everyday scenes for complex scene analysis.

Ego-centric Image Data: Complementing the COCO dataset, we sample 1500 samples from an Ego-centric image data source which provides a unique, first-person perspective. This dataset is crucial for understanding human-centric activities and behaviors, offering an immersive insight into the user's daily visual experiences.

4.2 Evaluation Experiments

Case Study: We consider a case comparing two different MM-LLMs (following the AnyMAL architecture (Moon et al., 2023)), each trained with a different base vision encoder (ViT-L vs. ViT-BigG (Radford et al., 2021)).

We evaluate the performance of the Multimodal Evaluator (MAEV) through two distinct tasks:

Task 1. Pointwise evaluation based on human raters agreement: This task focuses on the evaluation of single responses from the MM-LLM, com-

Base Evaluator	k-shot Prompts	Ego		Coco	
		F1	Acc.	F1	Acc.
Llama2	$k = 0$	0.698	64.96%	0.709	**68.52%**
	1	0.778	71.28%	0.720	61.18%
	2	0.741	68.59%	0.724	63.59%
	sampling	**0.792**	**72.49%**	**0.730**	64.85%
GPT4	0	0.756	67.24%	N/A	
	1	0.735	65.17%		
	2	0.728	64.48%		
	sampling	**0.846**	**76.92%**		

Table 1: **Performance of different base Evaluator LLMs** as measured by agreement rate with human evaluation results, across k-shot prompts on a candidate MM-LLM. Human annotators review the same data point multiple times and the most common score out of 5 trials is chosen as ground truth. The MAEV evaluators are then compared to this ground truth to measure accuracy and F1 score. MAEV with the sampling method exhibits the highest correlation with human evaluation.

Model	Description	MAEV score	Human score
Model A	Base Model	57.3%	56.3%
Model B	Larger vision encoder	58.3%	59.6%

Table 2: **Case Study: Accuracy measurement (%) of a candidate MM-LLM model** as scored by humans and MAEV, on the same test set. Accuracy as measured by MAEV closely tracks the human-measured accuracy, in both the absolute score ($\Delta < 1.5\%$) as well as their relative ranks (Model B is better than Model A).

paring them to the established ground truth answers. The ground truth is determined by having human raters evaluate an MM-LLM's response using the same rubric. We employ a multi-review system of 5, meaning each data point is independently scored 5 times by different human annotators. The most frequently given score is selected as the ground truth. With this ground truth as a reference, we then assess the Evaluator LLM's capability to score a response, using Accuracy and F1 score as our evaluation metrics. This approach provides a comprehensive analysis of how closely each response aligns with the anticipated result.

Task 2. Pointwise evaluation for selecting different (unseen) MM-LLM models: This task focuses on a real-world scenario where multiple MM-LLM candidates under development are being compared. We compare the scores predicted by the MAEV evaluator and human raters to determine whether there is agreement on which model performs better. This method allows for understanding comparative performance of multiple MM-LLMs.

4.3 Results

Table 1 shows the task 1 performance on the two data sets – MAEV COCO and MAEV Ego.

Shot sampling and Multi-trial aggregation improve the accuracy and robustness of evaluation: In all three instances of evaluation, sampling a variety of examples in the prompt shows the largest gains in performance. While increasing the number of examples in the context shows improvements, this can be encumbered by the maximum input context length size of the models.

Additionally, in a multimodal setting, the range of scenarios and inquiries presented to an image can be diverse. Thus, providing a variety of examples and aggregating the scores over multiple trials exposes the evaluator LLM to make a more balanced and generalizable assessment as indicated by the highest f1 scores in their task.

Rich text captions for images are important for multimodal evaluation: While MAEV-COCO comes with a diverse set of images and questions, the image captions can be brief and succinct. Thus leaving out crucial image details that might otherwise help determine if the model is hallucinating an object. To address this, we collected the MAEV-Ego dataset with descriptive image captions and informative reference answers, that are reflected in its higher F1 and accuracy scores.

LLAMA2 performance is comparable to GPT4 for Multimodal evaluations: Unlike past studies in text only evaluation, the results here indicate a narrower gap between LLAMA2 and GPT in MM-LLM evaluation. This could be potentially due to the high quality reference answers and image captions within MAEV-COCO and MAEV-Ego that we provide as part of the evaluation dataset. Consequently, the gaps in performance for both the models is in the image capabilities that can't be proxied through image captions.

MAEV can be used for model selection: In Table 2, comparing two different MM-LLMs scored by MAEV and human raters indicates agreement in the higher performing models. The results here show promise in using the framework and dataset for model selection.

Conclusions. Our research makes significantly advances in the evaluation of MM-LLMs. The MAEV framework and the accompanying datasets provide a robust tool for assessing the performance of these models, with promising implications for future model development and selection.

5 Limitations

While our research makes significant strides in the evaluation of multimodal large language models (MM-LLMs), it is not without its limitations, which we acknowledge as opportunities for future work.

Dataset Size and Diversity: Although we develop two new datasets for MM-LLM evaluation, the size and diversity of these datasets may still be limited. Future work can focus on expanding these datasets, both in terms of volume and variety of data, to provide a more comprehensive evaluation of MM-LLMs.

Model Performance: Our results indicate that the performance of the models is largely dependent on the quality of reference answers and image captions. However, the performance gaps for both models are found in their image capabilities, which cannot be proxied through image captions. Future research can explore ways to improve the image capabilities of these models.

Evaluation Framework: While our MAEV framework provides a robust tool for assessing the performance of MM-LLMs, it may not cover all possible aspects of model performance. Future work can aim to refine and expand the evaluation framework to include other important aspects of model performance.

Generalizability: Our research is based on two specific MM-LLMs. While we believe our findings are applicable to other similar models, the generalizability of our results to all MM-LLMs is not guaranteed. Future studies can aim to validate our findings across a wider range of MM-LLMs.

Human Evaluation: Our pointwise evaluation results indicate agreement between MAEV and human raters. However, human evaluation is inherently subjective and may not always be consistent. Future work can explore ways to improve the reliability and consistency of human evaluation.

In conclusion, while our research has its limitations, we believe it provides a solid foundation for future work in the evaluation of MM-LLMs. We look forward to seeing how our work can be built upon and enhanced in future studies

6 Ethics and Broader Impacts

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ACL Code of Ethics* and honor the code of conduct. We state the ethical considerations and the potential impact to the community as follows.

Dataset. Our main focus for human annotation is on the captions and responses given in image and text pairs, which are annotated by 20 workers with linguistic expertise. We provide detailed guidance and examples for the annotations and encourage diversity among the annotators, without any restrictions on their background as long as they are proficient in English and have domain expertise.

The data annotation and evaluation task was outsourced to a vendor specializing in NLP annotations, where the annotators are full-time employees. The annotators were given clear instructions, including a clear escalation path for reporting any sensitive topics that may arise in the seed images.

Techniques. We evaluate the performance of our constructed dataset using state-of-the-art pretrained language models, and adapt them to fit our specific tasks. Since our dataset is designed to predict the accuracy of model responses, we do not expect it to generate harmful outputs that could negatively impact vulnerable groups.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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

A Supplementary Materials

Hyperparameter	Value Range	Search Type	Final value
Temperature	[0, 1]	Random	Default API value
Top_p	[0, 1]	Random	0.2
Accuracy Score Threshold	(0,2)	Grid	1.25

Table 3: Hyperparameter Search Table

```
f"

[Question]

{question}

[Image Caption]
{caption}
[Reference Answer]
{reference}
[Assistant Answer]
{assistant}

[Instruction]
You are to evaluate the Answers provided by a multimodal AI assistant above.
Score the responses based on the following criteria and only the following
criteria:

[Evaluation Rubric]
**Accuracy**: Did the assistant's response correctly reference the main
object in the *caption* and *reference* without adding non-existent details?
Score 0: The response contains information that is not present in the caption
or reference answer.
Score 2: The response contains the main object in the caption or reference
answer and only includes accurate details.

[Your Evaluation] Respond in the following json format
'Evaluation Evidence': <Your analysis>
'Assistant Score': <Your score>

"

example_1 = random.sample(example_accuracy_0, 1)
example_2 = random.sample(example_accuracy_2, 1)

prompt += f"\n\n[Example]\n{example_1[0]}\n\n"
prompt += f"\n\n[Example]\n{example_2[0]}"
```

Figure 3: Prompt Template for the Evaluator LLM.

```
[Example 1]
[Question]
"Write an Instagram caption for this image."
[Image Caption]
"A busy urban street scene at night, highlighted by neon signs and a diverse
crowd."
[Reference Answer]
"City lights and starry nights: the urban jungle comes to life! 🌃✨
#CityThatNeverSleeps"
[Assistant Answer]
"Tranquility in nature's embrace, away from the urban race. 🌲🏞 #IntoTheWild"
[Output]
Evaluation Evidence: The assistant's caption contrasts with the urban scene,
inaccurately portraying a nature setting.
Assistant Score: "0"


[Example 2]
[Question]
"Describe the emotion conveyed in the painting."
[Image Caption]
"A somber mood is depicted through deep blue hues."
[Reference Answer]
"The painting showcases a profound sense of melancholy, with deep blue hues
dominating the scene, reflecting the subject's inner turmoil."
[Assistant Answer]
"The painting, with its bright reds and yellows, depicts a sense of intense passion
and joy."
[Output]
Evaluation Evidence: Incorrect colors and emotions cited, conflicting with the
painting's described mood.
Assistant Score: "0"
```

Figure 4: Examples used in Shot Sampling and Multi Trial aggregation where the accuracy score is 0

```
[Example 1]
[Question]
"What's the mood in the picture?"
[Image Caption]
"A bright sunny day at the park with people picnicking and playing around."
[Reference Answer]
"The picture portrays a cheerful mood, highlighted by people engaging in various
recreational activities in the park under the bright sun."
[Assistant Answer]
"The photo captures a joyful day, sunshine illuminating the park, while individuals
enjoy picnics and games in the warmth."
[Output]
Evaluation Evidence: The response correctly mirrors the cheerful atmosphere
depicted, with no inaccuracies.
Assistant Score: "2"

[Example 7]
[Question]
"What does the image depict?"
[Image Caption]
"An old, rustic bridge over a calm river, surrounded by autumn-colored trees."
[Reference Answer]
"The image depicts a quaint, old bridge crossing over a serene river, with trees
showcasing a spectrum of autumn hues, creating a peaceful scenery."
[Assistant Answer]
"Displayed is a rustic bridge, aged by time, spanning a tranquil river, while trees
dressed in autumn's palette complement the serene vista."
[Output]
Evaluation Evidence: The assistant's description aligns with the image, accurately
detailing the bridge, river, and trees.
Assistant Score: "2"
```

Figure 5: Examples used in Shot Sampling and Multi Trial aggregation where the accuracy score is 2