

## APPENDIX

### A. GPT PROMPTS

#### A.1 Filtering out irrelevant questions

```
system_prompt = "You are an expert in finding the relevance of a question and its corresponding answer with respect to a particular domain. Your task is to find the relevance of a question answer pair with respect to the domain mentioned. Here's how you can accomplish the task.
-----
- See whether the question is relevant to the domain in consideration.
- See whether the answer tries to provide a solution to the question
.
- Evaluate the question answer pair with respect to the domain mentioned.
- Rate the relevance of the following question-answer pairs on a scale of 1-10, with 1 being least relevant and 10 being most relevant.
Please generate the response in the form of a Python integer
DO NOT PROVIDE ANY OTHER TEXT OR EXPLANATION"

user_prompt = "Please evaluate the following domain-based question-answer pair:
domain:{domain}
question:{question}
answer:{answer}"
```

Listing 3: Prompt used to filter irrelevant QA pairs

### A.2 Prompts for Evaluation

#### A.2.1 Correctness

```
system_prompt = "You are an AI Evaluation chatbot helpful in evaluating the correctness of generative outputs for visual content-based question-answer pairs.
Your task is to evaluate the predicted answer and determine if it answers the question correctly. Here's how you can accomplish the task:
-----
##INSTRUCTIONS:
Provide your evaluation only as a score where the score is an integer value.
Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the score in INTEGER, not STRING.
For example, your response should look like this: {'score': INTEGER}.
Possible values of score are : [0,1,2]
0 when predicted answer is incorrect or not answered.
1 when predicted answer is partially correct or makes assumptions.
```

```
2 when predicted answer is completely correct with concise, useful explanations. "
```

```
user_prompt = "Please evaluate the following video-based question-answer pair:
Question: {question}
Predicted Answer: {pred}"
```

Listing 4: Prompt used for evaluating correctness

#### A.2.2 Coherence

```
system_prompt = "You are a strict AI Evaluation chatbot helpful in evaluating the coherence of generative outputs for video-based question-answer pairs.
Your task is to strictly evaluate the predicted answer and determine if it is coherent and easy to understand. Here's how you can accomplish the task:
-----
##INSTRUCTIONS:
Provide your evaluation only as a score where the score is an integer value.
Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the score in INTEGER, not STRING.
For example, your response should look like this: {'score': INTEGER}.
Possible values of score are : [0,1]
0 when predicted answer is long and not easy to follow.
1 when predicted answer has concise explanations and easy to follow."

user_prompt = "Please evaluate the following video-based question-answer pair:
Question: {question}
Predicted Answer: {pred}"
```

Listing 5: Prompt used for evaluating coherence

#### A.2.3 Visual grounding

```
system_prompt = "You are a strict AI Evaluation chatbot helpful in evaluating the contextual understanding of generative outputs for video-based question-answer pairs.
Your task is to strictly evaluate the predicted answer and determine if it is grounded in the visual content provided as image/video. Here's how you can accomplish the task:
-----
##INSTRUCTIONS:
Provide your evaluation only as a score where the score is an integer value.
Please generate the response in the form of a Python dictionary string
```

```

        with keys 'score', where its
        value is the score in INTEGER, not
        STRING.
    For example, your response should look
    like this: {'score': INTEGER}.
    Possible values of score are : [0,1]
    0 when predicted answer has no
    reference to the image/video.
    1 when predicted answer has reference
    to the image/video. "

user_prompt = "Please evaluate the
following video-based question-answer
pair:
Question: {question}
Predicted Answer: {pred}"

```

Listing 6: Prompt used for evaluating visual understanding

#### A.2.4 Relevance

```

system_prompt = "You are a strict AI
Evaluation chatbot helpful in
evaluating the completeness and
relevance of generative outputs for
video-based question-answer pairs.
Your task is to strictly evaluate the
predicted answer and determine if
it is complete and relevant to the
question. Here's how you can
accomplish the task:
-----
##INSTRUCTIONS:
Provide your evaluation only as a
score where the score is an
integer value.
Please generate the response in the
form of a Python dictionary string
with keys 'score', where its
value is the score in INTEGER, not
STRING.
For example, your response should look
like this: {'score': INTEGER}.
Possible values of score are : [0,1]
0 when predicted answer is irrelevant
or somewhat deviates from the
question.
1 when predicted answer is completely
relevant and precise. "

user_prompt = "Please evaluate the
following video-based question-answer
pair:
Question: {question}
Predicted Answer: {pred}"

```

Listing 7: Prompt used for evaluating relevance

## B. LIMITATIONS

We would like to extend this benchmarking to many other models like Gemini [24] and GPT-4o. It will also be nice to distill such large models to small scale models for lower compute needs and lower latency.

We experimented with maths and science questions. It will be nice to extend this to other domains.

Lastly, we experimented with English QA pairs only. We would surely like to extend this to more languages.

## C. ETHICS STATEMENT

All the models used in this work are publicly available on Huggingface and free for research.

We utilized publicly accessible LearningQ dataset from <https://github.com/AngusGLChen/LearningQ>. These resources were used as per their intended use policies.

Just like other generative models, our models can potentially generate biased, offensive or otherwise harmful content. Hence, care should be taken to apply appropriate filters when integrating with real world systems. That said, we did not observe such cases during our experimentation.