

# Self-Rambling Attack against Large Vision-Language Models Serving

## –Supplementary Material–

### I. OVERVIEW

Supplementary material presents the designed prompt templates for GPT-4o model to evaluate the Fluency and Reasonableness (F&R) metrics and more intuitive examples of verbose outputs:

#### A. Designed Prompt Templates to Evaluate the F&R metrics

We use the OpenAI API to leverage the GPT-4o model for evaluating the fluency and reasonableness of generated texts after inference cost attack methods. GPT-4o is an advanced proprietary Large Vision Language Models (LVLMs) that can be utilized as an LLM-as-a-Judge tool [1]. Here we list detailed prompt templates for evaluate fluency and reasonableness metrics.

```

1 # Fluency Scoring Prompt Template
2 response = openai.ChatCompletion.create(
3     model="gpt-4o",
4     messages=[
5         {
6             "role": "system",
7             "content": (
8                 "You are a senior multimodal
9                 evaluation specialist. "
10                "Your task is to assess the
11                linguistic fluency of a
12                given text as a
13                description of a provided
14                image. "
15                "You must evaluate along
16                these four
17                dimensions:\n\n"
18                "1. Grammatical Correctness:
19                Are syntax and grammar
20                accurate and
21                error-free?\n\n"
22                "2. Lexical Richness: Is the
23                vocabulary varied,
24                precise, and
25                appropriate?\n\n"
26                "3. Cohesive Flow: Does the
27                text read smoothly with
28                clear transitions and
29                logical progression?\n\n"
30                "4. Clarity of Expression: Is
31                meaning conveyed
32                unambiguously and
33                concisely?\n\n"
34                "After your analysis, provide
35                a single numeric score
36                between 1.0 and 5.0 (one
37                decimal place) reflecting
38                overall fluency, "

```

```

        "plus a brief explanation
        (1-2 sentences) stating
        which dimension(s) most
        influenced your score. "
        "Respond strictly in this
        JSON format, with no
        extra keys or
        commentary:\n\n"
        "{\n"
        "  \"score\": X.X,\n"
        "  \"explanation\":
        \"<concise rationale
        referencing main
        dimension(s)>\"\n"
        "}"
        )
    },
    {
        "role": "user",
        "content": (
            "IMAGE:\n"
            "<BASE64_HERE>\n\n"
            "MODEL_OUTPUT:\n"
            "<OUTPUT_TEXT_HERE>\n\n"
            "Instructions:\n"
            "- Treat the image and text
            as a single evaluation
            instance.\n"
            "- Use temperature=0.0 for
            deterministic output.\n"
            "- Ensure `score` has exactly
            one digit after the
            decimal point.\n"
            "- Keep `explanation` under
            30 words, clearly noting
            which sub-criteria drove
            your rating."
        )
    }
    ],
    temperature=0.0,
    max_tokens=2048
)

```

```

1 # Reasonableness Scoring Prompt Template
2 response = openai.ChatCompletion.create(
3     model="gpt-4o",
4     messages=[
5         {
6             "role": "system",
7             "content": (
8                 "You are a senior multimodal
9                 evaluation specialist. "
10                "Your task is to assess how
11                reasonable a given text
12                is as a description of a
13                provided image. "

```

```

10     "You must evaluate along
11     these four
12     dimensions:\n\n"
13     "1. Information Relevance:
14     Does the text address the
15     main subjects, actions,
16     and context shown in the
17     image?\n"
18     "2. Visual Consistency: Are
19     the objects, colors,
20     spatial relationships,
21     and attributes described
22     actually present and
23     correctly represented?\n"
24     "3. Logical Coherence: Is the
25     narrative
26     commonsense-correct, free
27     of internal
28     contradictions, and
29     consistent with
30     real-world knowledge?\n"
31     "4. Detail Completeness: Are
32     key elements and details
33     covered without major
34     omissions, while avoiding
35     hallucinated or
36     irrelevant details?\n\n"
37     "After your analysis, provide
38     a single numeric score
39     between 1.0 and 5.0 (one
40     decimal place) reflecting
41     overall reasonableness, "
42     "plus a brief explanation
43     (1-2 sentences) stating
44     which dimension(s) most
45     influenced your score. "
46     "Respond strictly in this
47     JSON format, with no
48     extra keys or
49     commentary:\n\n"
50     "{\n"
51     "  \"score\": X.X,\n"
52     "  \"explanation\":\n"
53     "    \"<concise rationale\n"
54     "    referencing main\n"
55     "    dimension(s)>\"\n"
56     "}"
57     "\n"
58     },
59     {
60       "role": "user",
61       "content": (
62         "IMAGE:\n"
63         "<BASE64_HERE>\n\n"
64         "MODEL_OUTPUT:\n"
65         "<OUTPUT_TEXT_HERE>\n\n"
66         "Instructions:\n"
67         "- Treat the image and text
68         as a single evaluation
69         instance.\n"
70         "- Use temperature=0.0 for
71         deterministic output.\n"
72         "- Ensure `score` has exactly
73         one digit after the
74         decimal point.\n"
75         "- Keep `explanation` under
76         30 words, clearly noting
77         which sub-criteria drove
78         your rating."
79       )
80     }
81   ],
82   temperature=0.0,
83   max_tokens=2048
84 )

```

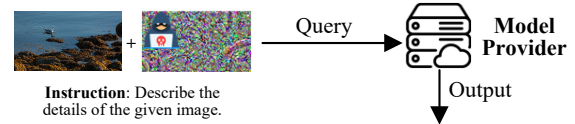
```

36   )
37   }
38   ],
39   temperature=0.0,
40   max_tokens=2048
41 )

```

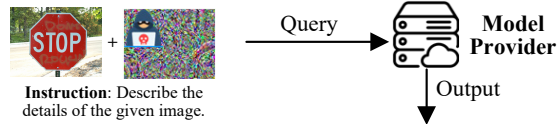
## B. More Intuitive Examples

Here we illustrate more output texts after our Self-Rambling (S-R) attack on BLIP-2 2.7B. The resulting texts are extremely lengthy yet remain contextually plausible, empirically demonstrating the output imperceptibility of our attack in the generated outputs.



**Attacked Output:** On the left of the image there is a small bird shaped object that has a piece of wood sticking out of it. The bird has a piece of wood and a broken leg sticking out of it. The bird has been floating in the waters in the picture since the whole picture was taken. The bird has not been in the water for a long time. The piece of wood seems to be in the middle of the water. There is something black in the middle of the water that is also in the picture. It looks like it could be the bird's foot. The bird in the picture is struggling to get to its feet. The pole sticking out of the bird looks as if it is submerged into the water. This is because the water seems to be calm but it looks as if it is stirred up from the sea life or by something like a wave. The woods would be for protection. The wood would not be soft because it is from an old time piece and therefore would be hard. The piece of wood on the bird's leg would not be like a wooden board because it would have been broken by the current. This would give it more protection. The pole would be in a swimming motion because it is hard. This means that it would move away from the bird instead of toward the bird. The plank would be what it would be floating on. this is a close up of a blue duck standing in a puddle with a toy boat beside it on a piece of brown seaweed, swimming in the clear green water next to a rusted old rusty barrel on the left edge of the photo. You can see the rusted iron drain in the picture and the shiny metal buoy to the right of the duck in the rusted old rusty barrel that has some seaweed growing on the top. Please don't brighten it. Let's keep it as natural as possible. You need to get rid of that sky and be more precise on the way you change the focal distance and the way you crop it, and make it more close up like it is in this photo. The sky and the surrounding area is fine as long as you make it sharp and close up. Also, get rid of that water surface if possible. The blue duck is the star of the show in this photo. The other details should not be too close or bright or they will steal the focus from the blue duck and make it too fuzzy. This photo is a photo of a sea level pond where a blue duck is standing in the water. This blue duck is standing on a piece of sea weed. **[473 Tokens]**

**Original Output:** there is a seagull in the water and a bunch of seaweed is lying on the surface of the water **[14 Tokens]**



**Attacked Output:** The photo is a traffic stop sign on a pole. The image is posted on a wood panel of some type. The sign is facing the camera and the ground behind the sign is wet and soggy. I see a barricade in front of the sign which says in someones handwriting "dont stop" on it. A fence is in the foreground with two posts. Next is a tree with leaves and cones. On the other side of the trees is a curb.the picture is a stop sign on the side of the road. The sign reads no right turn. When looking at the picture you have to imagine you are driving. It is off the main road to the right. There is a tree by the sign that is keeping it from blowing away. When you think about it, you may notice the lettering of the sign is hard to read because it has been covered with green plastic. It appears the paint has already been removed from the lettering. It has also been blocked out with yellow plastic. At the top of the sign the red letters say stop. There is a white line on the sign and that white line is the line of sight to the road. You can see the road across the street is also empty. The tree, house and trash can are all in the background. There are also some cars in the background. **[209 Tokens]**

**Original Output:** On the underside of the sign is written 'do not paste.' Below is written in black marker: "do not paste." **[16 Tokens]**

Fig. 1: Intuitive cases illustration.

## REFERENCES

- [1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.

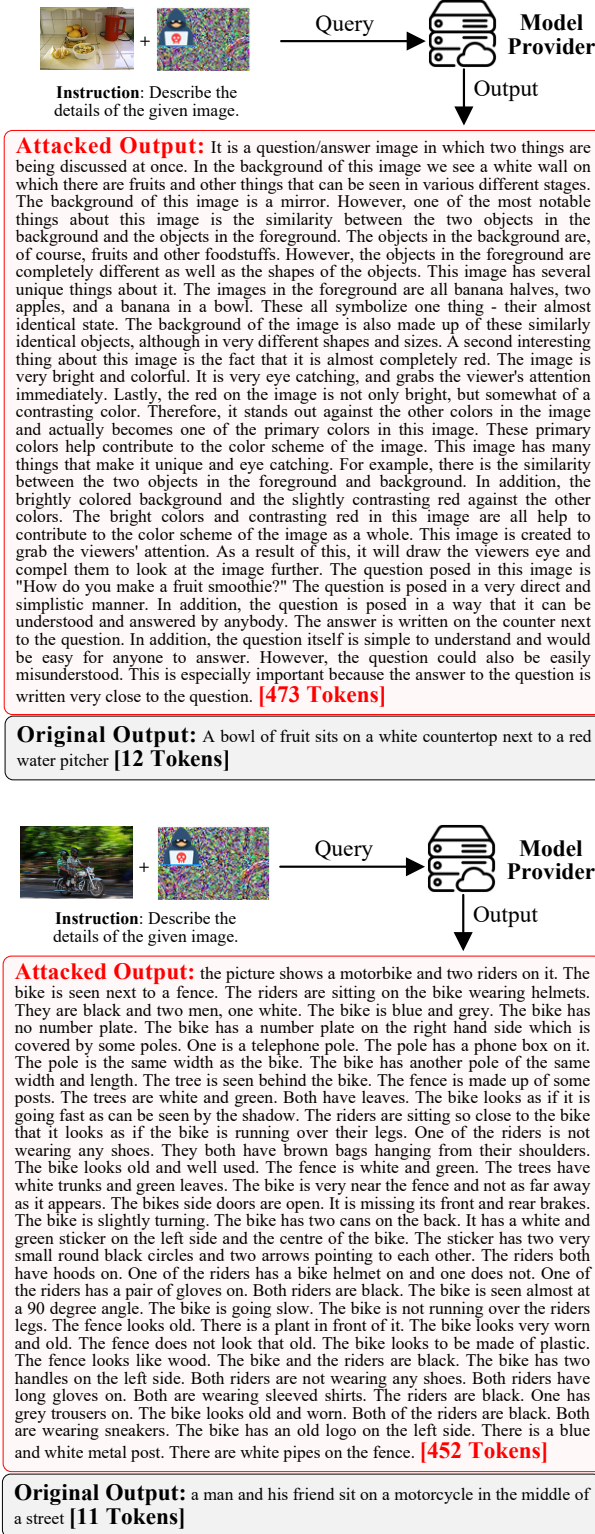


Fig. 2: Intuitive cases illustration.