

1 A Prompt Templates

Prompt for cold-start data synthesis

Answer the question and provide your reasoning process, including the following:

1. Simulate image reasoning: Treat the image caption as an image. Simulate reasoning by imagining you are looking at the image, and act as if you can see it. However, avoid visualization as a step in the reasoning process.
2. Direct visual language: Frame observations as if you are directly viewing the image (e.g., "The image shows..."). Avoid reasoning through image caption or description.
3. Forbidden phrases: Avoid phrases like "based on the caption", "based on the description", "visualizing the image", "the OCR results given".
4. You need output the final OCR in the end. Question: output the reasoning process to get the result as below.

Question: {question}

Image Content: {caption}.

Character info: {character information}.

Prompt for zero-shot evaluation

You are a highly specialized OCR model that processes images containing text and numbers. Your task is to perform character-level OCR and categorize each character based on its clarity. If you are unable to recognize a character or number, you must explicitly refuse to recognize it. For each other readable character, categorize it as either "clear" or "not clear enough". Finally, generate an output in the following format:

- Clear char-level OCR: List characters that are clearly recognize, separated by spaces.
- Not clear enough char-level OCR: List characters that are recognized but not clear enough, separated by spaces.
- Clear number: Count of numbers that are clearly recognized.
- Not clear enough number: Count of numbers that are recognized but not clear enough.
- Final OCR: Compile all recognized characters and numbers into a single output, maintaining their original order.

Output Format:

```
{  
  "clear char-level OCR": "0 1 5 S t i t i",  
  "not clear enough char-level OCR": "2",  
  "clear number": 8,  
  "not clear enough number": 1,  
  "final OCR": "2015 S t i t i"  
}
```

4 B Limitations and Broader Impacts

Limitations. In this section, we provide an analysis of the limitations of the method proposed in this paper. Regarding computational overhead, the GRPO method based on multi-objective reward functions inevitably brings additional computational costs, and the fine-grained character-level OCR hallucination benchmark KIE-HVQA relies on manual annotation. In terms of application scenarios, OCR hallucinations are not always harmful. This paper focuses on scenarios where OCR hallucinations are detrimental, such as ID card recognition and receipt recognition. How to balance the impact of OCR hallucinations on the final results in different scenarios will be considered in future work.

Broader Impacts. Our work primarily focuses on addressing the widespread OCR hallucination issue present in existing MLLMs. A potential positive societal impact is assisting OCR systems or products in improving OCR accuracy, particularly in key information extraction (KIE) tasks, thereby enhancing user experience. A potential negative societal impact is that the processing of document images might raise concerns about privacy and security among users. A feasible solution is to fully inform users about the detailed processing procedures of the model and offer an option not to process images containing private information.

20 C Visualization

21 In this section, we provide additional examples from the KIE-HVQA benchmark, along with corre-
 22 sponding questions and annotated answers (Q & A), as well as the zero-shot results of the Qwen2.5-
 23 VL-7B model.

Q: What is the value of Agama?

A: ["clear Char-level OCR: '1 s l a m',
 "not clear enough Char-level OCR: "",
 "clear number": 5,
 "not clear enough number": 0,
 "Final OCR: 'Islam'"]



WORD DEFINITION

NIK

507090224042000

Nama: Ningsih
 Tempat/Tgl. Lahir: Yogyakarta/1980-05-05
 Jenis kelamin: Perempuan
 Alamat: J. Matelidoro No. 2
 RT/RW: 001/010
 Kel/Desa: Kertanegara
 Kecamatan: Kertanegara
 Kabupaten: Bantul
 Provinsi: Yogyakarta
 Agama: Islam
 Status Perkawinan: Belum Kawin
 Pekerjaan: Mahasiswa
 Kewarganegaraan: Indonesia
 Berlaku Hingga: 2025-05-05

D Sign

Q: What is the value for 'LOT.#'?

A: ["clear Char-level OCR: '9 1 2 5 9',
 "not clear enough Char-level OCR: "",
 "clear number": 5,
 "not clear enough number": 0,
 "Final OCR: '91259'"]

A: ["clear Char-level OCR: '0 1 2 5 9',
 "not clear enough Char-level OCR: "",
 "clear number": 6,
 "not clear enough number": 0,
 "Final OCR: '01259'"]



Q: What is the value for Total Fat of per 100g/ml?

A: ["clear Char-level OCR: '1.5 g',
 "not clear enough Char-level OCR: "",
 "clear number": 4,
 "not clear enough number": 0,
 "Final OCR: '1.5g'"]

A: ["clear Char-level OCR: '5 g',
 "not clear enough Char-level OCR: "",
 "clear number": 3,
 "not clear enough number": 0,
 "Final OCR: '5g'"]



Q: What is the value of Agama?

A: ["clear Char-level OCR: 'S90 23456',
 "not clear enough Char-level OCR: "",
 "clear number": 8,
 "not clear enough number": 0,
 "Final OCR: 'S90 23456'"]

A: ["clear Char-level OCR: 'S 9 0 3 4 5 6',
 "not clear enough Char-level OCR: "",
 "clear number": 7,
 "not clear enough number": 0,
 "Final OCR: 'S90 3456'"]



Q: What was the cost of the Bibim Octopu?

A: ["clear Char-level OCR: '1 9 . 9',
 "not clear enough Char-level OCR: "",
 "clear number": 4,
 "not clear enough number": 0,
 "Final OCR: '19.9'"]

A: ["clear Char-level OCR: '\$ 1 . 9',
 "not clear enough Char-level OCR: '3',
 "clear number": 4,
 "not clear enough number": 1,
 "Final OCR: '\$13.9'"]



Q: What is the value of Agama?

A: ["clear Char-level OCR: '2 1 . 3 0',
 "not clear enough Char-level OCR: "",
 "clear number": 5,
 "not clear enough number": 0,
 "Final OCR: '21.30'"]

A: ["clear Char-level OCR: '2 . 3 0',
 "not clear enough Char-level OCR: "",
 "clear number": 4,
 "not clear enough number": 0,
 "Final OCR: '2.30'"]



Figure 1: Some cases in our KIE-HVQA benchmark.