

Supplementary Material of MME-Finance: A Multimodal Finance Benchmark for Expert-level Understanding and Reasoning

Ziliang Gan¹, Dong Zhang¹, Haohan Li¹, Yang Wu¹, Xueyuan Lin^{4,1} Ji Liu¹, Haipang Wu¹,

Chaoyou Fu², Zenglin Xu³, Rongjunchen Zhang^{1,♦}, Yong Dai¹

¹HiThink Research, ²Nanjing University, ³Fudan University,

⁴The Hong Kong University of Science and Technology (Guangzhou)

♦ denotes corresponding author, zhangrongjunchen@myhexin.com.

Abstract

In this supplementary material, we provide additional details regarding the proposed MME-Finance. Section 1, Section 2, and Section 3 provide detailed descriptions of the content of the open-end version, binary-choice version, and multi-turn version, respectively. These versions include more detailed statistical data, experimental results, and illustrative examples.

CCS Concepts

- Computing methodologies → Natural language generation.

Keywords

Benchmark, Finance, Multimodal Large Language Model

1 Open-end Version

1.1 Statistic

Statistic of the English version. Table 1 shows the statistic of open-ended English version of MME-Finance, which contains 1,171 image-question-answer pairs. The number of samples per task varies from 18 to 229, the “Spatial Awareness” task contains the most, and the “Reason Explanation” contains the fewest. The distribution of 6 types and 4 styles are shown in the Figure 1(a) and Figure 1(b), respectively. Statistics charts have the highest proportion, while mixed charts have the lowest. As for styles, computer screenshots have the largest number.

Statistic of the Chinese version. As shown in Table 2, the open-ended Chinese MME-Finance contains 1,103 image-question-answer pairs and has the same task categories as the English version. The number of samples per task varies from 13 to 182, with the “OCR” task containing the most and “Reason Explanation” the fewest. The distribution of 6 types of image and 4 styles of image are shown in the Figure 2(a) and Figure 2(b), respectively. Statistical charts have the highest proportion, while mixed charts have the lowest. The style of mobile photography has the highest number.

1.2 Comparison of image distribution

To further analyze MME-Finance, we compared the visualization results of the images in MME-Finance with other chart benchmarks, such as ChatQA [1], ChatX [2], UniChat [3], and PlotQA [4]. 200 images were randomly selected for each benchmark, respectively. As shown in Figure 3, the distribution of images in MME-Finance is significantly different from other benchmarks, indicating that there is a large domain gap between MME-Finance and other benchmarks. This is because the images in our benchmark are mainly professional and financial, which is quite different from the charts

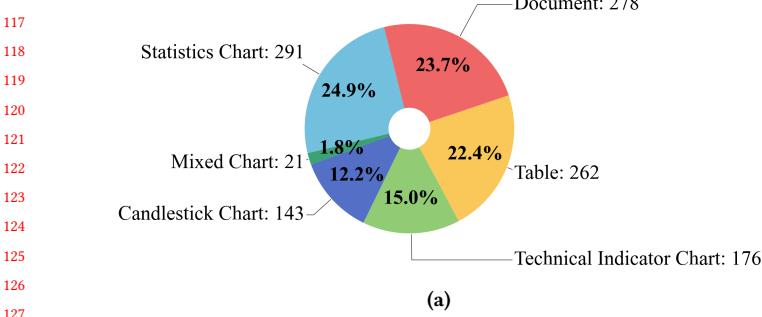
Table 1: Statistic of the number of samples in different capabilities and tasks in the English MME-Finance.

Statistic	Number
Perception	734
- Image Caption	164
- OCR	178
- Entity Recognition	163
- Spatial Awareness	229
Reasoning	175
- Accurate Numerical Calculation	133
- Estimated Numerical Calculation	42
Cognition	240
- Risk Warning	22
- Investment Advice	53
- Reason Explanation	18
- Financial Question Answer	147
Hallucination	22
- Not Applicable	22

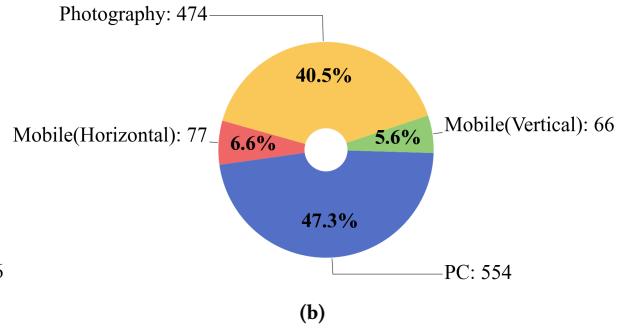
Table 2: Statistic of the number of samples in different capabilities and tasks of the Chinese MME-Finance.

Statistic	Number
Perception	640
- Image Caption	144
- OCR	182
- Entity Recognition	148
- Spatial Awareness	166
Reasoning	158
- Accurate Numerical Calculation	126
- Estimated Numerical Calculation	32
Cognition	285
- Risk Warning	37
- Investment Advice	91
- Reason Explanation	13
- Financial Question Answer	144
Hallucination	20
- Not Applicable	20

in other benchmarks. In addition, it can be seen that our data are clustered into several centers, which indicates that the images in MME-Finance are collected by category.

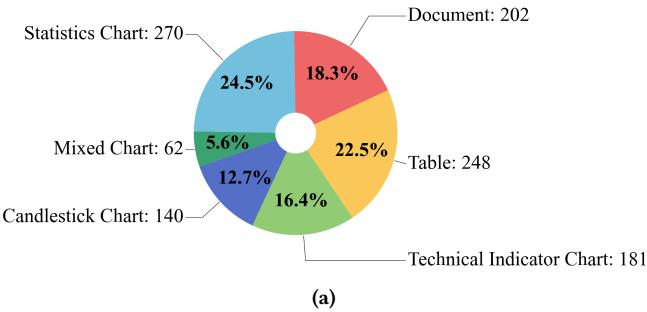


(a)

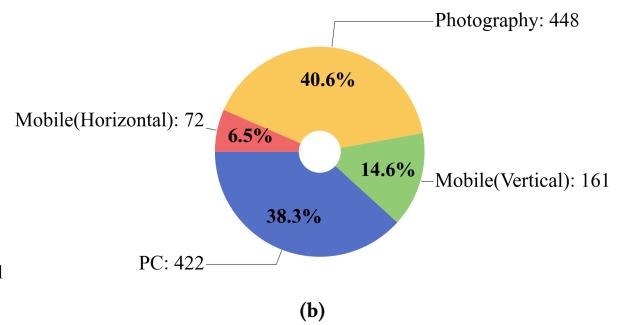


(b)

Figure 1: Distribution of different (a) types and (b) styles of images in English MME-Finance.



(a)



(b)

Figure 2: Distribution of different (a) types and (b) styles of images in Chinese MME-Finance.

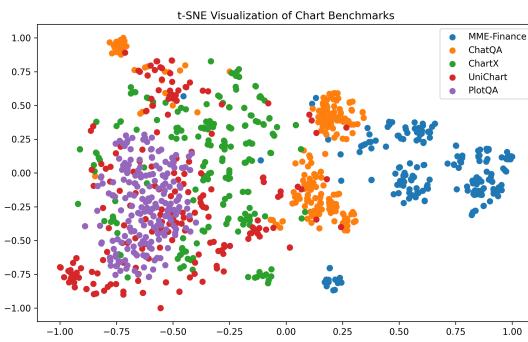


Figure 3: Visualization of several chart benchmarks.

1.3 Details of Annotation Pipeline

To make sure the quality of the annotation, the Chinese and English data were reviewed by two teams of financial experts in our company, each team specializing in the Chinese and North American financial markets, respectively. In addition, we have developed a comprehensive pipeline aimed at enhancing the quality and efficiency of annotation. For the objective questions, we divide the annotators into 3 groups, 2 groups for annotation, and 1 group for revision. The two annotation groups modify the pre-annotation, respectively, then the revision group checks whether the two annotations (for the same question) are consistent. We train annotators

to increase the consistency rate from 70% to 95% before the official annotation. Those inconsistent annotations will undergo further revision. For subjective questions, we first train the annotators until the BGE similarity of two annotations reaches at least 75%. In the official annotation phase, these annotators are divided into multi-groups with a group of 3 people, two of whom make revision of the pre-annotation, then they will discuss with another more senior expert to get the final answer.

1.4 Inference and Evaluation Pipeline

Figure 4 shows the inference and evaluation pipeline. During the inference phase, meticulously designed prompts are employed to guide and constrain the output formats of MLLMs, thereby ensuring a more standardized and reliable evaluation process. In the scoring phase, we utilize an LLM-based evaluation system to compare the prediction of the model against both the ground truth and the corresponding image, thereby assigning a comprehensive score.

1.5 Experimental Results

In this section, we report the evaluation results of all the models. **English Version.** As shown in the Table 3, proprietary MLLMs such as Gemini2.5Pro [5], GPT-4o [6] achieve better performance than other models. For open source MLLMs, InternVL3-78B [7] and Qwen2.5vl-72B [8] outperform other models. However, their performance remains slightly lower than that of the top-performing models. Table 4 presents the performance of various MLLMs from the perspective of image types and styles. It is notably that Gemini2.5Pro [5] achieves the best results across the vast majority of

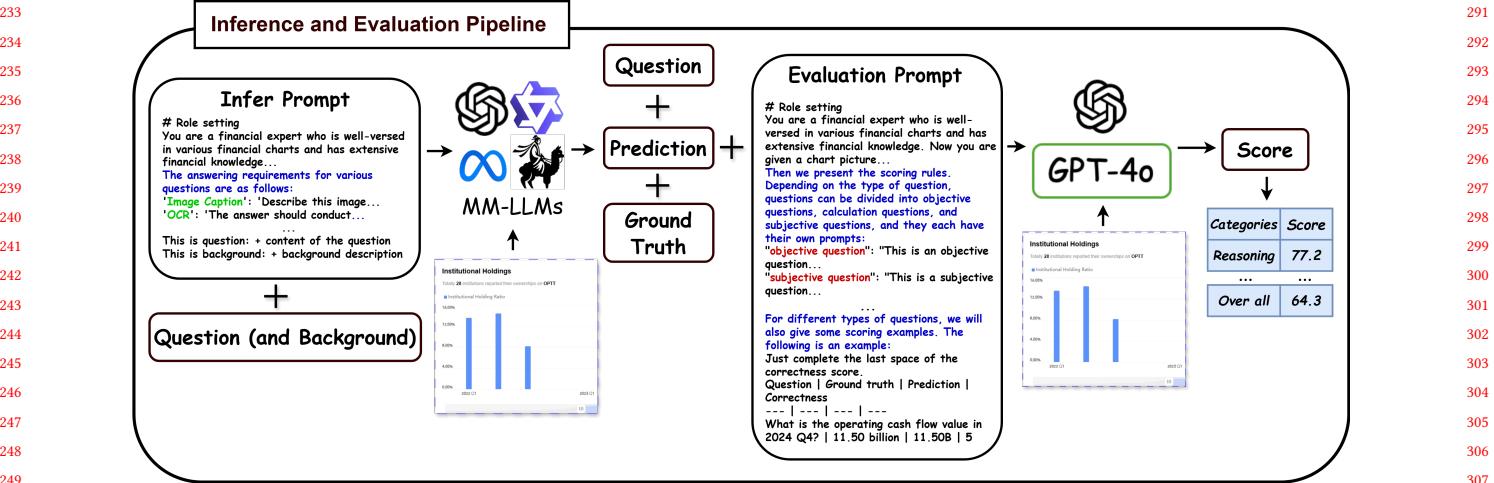


Figure 4: Inference and evaluation pipeline of MME-Finance. We first input the image, question and inference prompt into the MLLMs to obtain prediction. Then we feed the image, question, prediction, ground truth and evaluation prompt into GPT-4o evaluator to obtain scores. The inference and evaluation prompts are all designed individually for each task category.

image types and styles, highlighting its superior performance and strong generalization capability. It can be seen that most models exhibit poor performance on candlestick charts and technical indicator charts. This can be attributed to the specialized nature of these charts, which require domain-specific knowledge that current MLLMs struggle to interpret. Regarding image styles, most MLLMs exhibit suboptimal performance when applied to mobile photographs, primarily attributed to the lower resolution of images captured by phones, which hampers the visibility of crucial details. Furthermore, oblique angles of some photos also lead to incomplete or extraneous visual information. ★ Given the prevalence of such image in real-world applications, it is imperative to enhance the proficiency of MLLMs in mobile photographs understanding.

We conduct a detailed analysis on low-score (0-1 score) responses of Claude 3.5 [9], and find that in the objective question 26.9% errors are extracting incorrect information from the image (the rich information in the image affects the model to extract the target information), 55.6% errors are perceiving distorted spatial relation (fine-grained perception and logical reasoning based on spatial cues are great challenges for models), and 17.5% misidentifying attributes (e.g. color, shape). In the subjective question, 10% low-score responses misunderstand the meaning of the question, around 50% of them do not make full use of the image or background information, and about 60% of them make a wrong conclusion based on the information.

Chinese Version. Table 5 shows the results of various open source MLLMs on Chinese MME-Finance on each task with Qwen2VL-72B [17] evaluator. Among these models, Gemini2.5Pro [5] achieves the best overall performance with 81.07% accuracy. The performance of Qwen2.5VL-72B is slightly lower, and the third is InternVL3-38B. The evaluation results for different types and styles of images are shown in Table 6. Gemini2.5Pro achieves the best results in almost all categories. It is obviously that most models have lower accuracy in the type of candlestick chart and mixed chart. In terms

of image styles, mobile photographs present the greatest challenge for most MLLMs. Overall, these results are similar to those of the English MME-Finance.

1.6 Results of NA Prompt For ALL Tasks

Table 7 and Table 8 show the performance of MLLMs with the prompt to allow “Not Applicable” response across all types of tasks. It is clear that most models have lower performance in the setting, which means that the hallucination problem is quite common in MLLMs. Taking Qwen2VL-72B for example, its overall accuracy drops from 65.69% to 62.97%. Except for the NA task, its accuracy decreases the most (5.03%) in the Entity Recognition (ER) task. In addition, it drops the most in the type of candlestick charts and the style of mobile vertical screenshot. Although some models have a high recall of the “Not Applicable” question, their overall accuracy is low. It shows that these models tend to answer “Not Applicable” for those unsure questions.

1.7 Hard Examples

In this section, we present some hard examples about the difficulty of mobile photos and the hallucination problems of MLLMs. As shown in Figure 5, the two questions have a similar content. When feeding the two images into the same model, the responses are different. For the picture taken with a mobile phone, the model mistakenly identifies decimal points as commas and the letter B as the number 8. In contrast, the model accurately identifies the corresponding elements in the computer screenshot. This indicates that the perception of mobile phone photos is a challenge for some MLLMs. Figure 6 illustrates an example of the hallucination problem. GPT-4o cannot recognize the initial increase trend, while Qwen2VL-72B cannot completely perceive the trends. These widespread hallucination problems have significantly hindered the application of MLLMs in real-world scenarios.

Table 3: Evaluation results on MME-Finance. Abbreviations adopted: IC for Image Caption; ER for Entity Recognition; SA for Spatial Awareness; FQA for Financial Question Answer; ANC for Accurate Numerical Calculation; ENC for Estimated Numerical Calculation; RW for Risking Warning; IA for Investment Advice; RE for Reason Explanation; NA for Not Applicable. The F1-score metric is adopted for binary-choice questions, and the MT overall metric is designed for multi-turn questions. The first, second, and third highest values are highlighted by orange, blue, and green backgrounds, respectively. All numbers are denoted in %.

Model	Overall	Perception			Reasoning			Cognition			NA	
		IC	OCR	ER	SA	ANC	ENC	RW	IA	RE		
Open source MLLMs												
Yi-VL-34B [10]	17.57	29.39	1.46	3.93	8.73	5.56	11.43	42.73	35.09	58.89	47.48	36.36
CogVLM2-19B [11]	46.32	67.32	61.24	35.83	16.59	44.51	33.33	59.09	52.83	31.11	58.64	93.64
InternVL2-2B [12]	37.42	59.63	46.97	21.23	18.52	28.27	19.05	59.09	50.94	60.00	51.70	33.63
InternVL2-4B [12]	47.69	67.44	58.88	33.74	18.95	55.49	30.48	68.18	54.34	64.44	60.95	59.09
InternVL2-8B [12]	53.58	71.71	68.43	38.28	25.33	62.86	37.14	72.73	60.75	76.67	63.13	61.82
InternVL2-76B [12]	61.62	83.17	77.64	47.60	30.31	70.08	41.90	75.45	66.42	76.67	72.24	79.09
InternVL3-2B [7]	53.07	71.22	73.48	41.84	28.91	60.15	31.43	57.27	51.32	67.78	53.74	69.09
InternVL3-8B [7]	65.69	74.39	84.27	60.00	44.63	76.99	43.33	65.45	56.60	73.33	69.53	76.36
InternVL3-9B [7]	65.89	77.93	84.72	60.74	43.84	75.34	40.48	69.09	56.98	70.00	70.34	68.18
InternVL3-14B [7]	69.02	79.51	84.72	64.05	49.52	80.90	46.19	66.36	57.36	71.11	72.24	82.73
InternVL3-38B [7]	67.75	79.51	86.07	56.81	46.29	82.11	47.14	69.09	56.98	74.44	70.88	87.27
InternVL3-78B [7]	71.24	79.51	89.66	62.45	49.61	88.27	49.05	75.45	58.49	74.44	74.97	89.09
LLaMA3.2-11B [13]	42.51	62.44	39.10	32.02	14.50	55.79	37.14	60.00	50.57	68.89	57.55	61.82
LLaMA3.2-90B [13]	48.76	64.27	46.74	41.27	25.85	55.64	22.86	63.64	61.13	64.44	65.58	81.82
LLaVA-Next-7B [14]	28.18	58.41	22.81	14.85	11.09	7.07	10.00	45.45	47.55	12.22	54.97	55.45
LLaVA-Next-13B [14]	31.37	62.68	25.39	22.58	10.31	12.63	9.05	47.27	40.00	12.22	59.46	78.18
MiniCPM2.6 [15]	51.65	71.22	63.71	37.67	24.37	55.64	21.43	72.73	58.87	66.67	66.80	77.27
Phi3-Vision [16]	46.69	69.88	57.64	28.34	18.08	47.52	34.76	65.45	58.11	68.89	57.41	100.0
Phi3.5-Vision [16]	38.99	67.56	33.03	18.90	20.52	32.33	19.52	67.27	55.85	72.22	54.42	93.64
Qwen2VL-2B [17]	44.42	62.07	66.07	28.47	20.09	44.36	23.33	53.63	44.53	58.89	53.47	68.18
Qwen2VL-7B [17]	44.44	62.19	64.49	26.50	19.04	45.56	27.62	57.27	48.30	58.89	54.97	68.18
Qwen2VL-72B [17]	65.69	82.56	87.52	55.46	27.16	83.76	40.95	78.18	65.66	77.78	75.37	90.91
Qwen2.5VL-3B [8]	57.85	63.54	82.47	49.08	30.83	70.53	46.67	61.82	52.83	68.89	60.14	90.91
Qwen2.5VL-7B [8]	62.00	72.93	82.47	49.08	30.83	70.53	46.67	61.82	52.83	68.89	60.14	90.91
Qwen2.5VL-32B [8]	65.41	75.61	86.29	49.33	42.79	81.95	45.71	71.82	59.24	70.00	69.93	87.27
Qwen2.5VL-72B [8]	68.20	75.73	87.64	61.60	40.00	84.21	58.09	72.73	58.49	72.22	75.37	87.27
Proprietary MLLMs												
Gemini1.5Pro [18]	61.84	82.20	80.22	48.59	23.14	78.20	50.95	76.36	69.43	75.56	70.75	80.91
Claude3.5-Sonnet [9]	63.91	87.80	63.70	54.23	35.46	72.33	60.00	80.91	72.83	82.22	73.33	95.45
GPT-4o-mini [6]	64.43	86.46	73.71	54.72	34.93	69.17	56.19	76.36	63.46	72.22	77.55	87.27
GPT-4o [6]	72.79	89.88	86.18	61.60	45.68	82.41	65.24	80.91	70.57	80.00	82.59	84.55
Gemini2.5Pro [5]	79.28	90.85	94.61	70.18	63.14	87.97	65.24	75.45	68.68	78.89	80.54	100.00

1.8 Inference and Evaluation Prompt

Figure 7 and Figure 8 show the detailed inference prompt and the evaluation prompt for English MME-Finance. The usage of these prompts can be found in our github code.

1.9 Definition and Example For Each Task

In this section, we provide a detailed definition of each task and present corresponding examples to help readers learn about these tasks.

Perception

- (1) **Image Caption:** Generate a textual description that accurately represents the content, context, and significant elements of an image.
- (2) **OCR:** Recognition of text, number in the image.
- (3) **Entity Recognition:** Recognition and understanding of visual elements(such as color, shape) in the image.
- (4) **Spatial Awareness:** Understand the position and spatial relationship of the elements in the image.

465 **Table 4: Evaluation results on English MME-Finance for different types and styles of images. Abbreviations adopted: Candle.** 523
 466 **for candlestick chart; Tech.** for technical indicator chart; **Stat.** for statistical chart; **Tab.** for table; **Doc.** for document; **Mixed** for 524
 467 **mixed chart; CS for computer screenshot; MP for mobile photograph; VS for vertical screenshot on Mobile; HS for horizontal** 525
 468 **screenshot on Mobile. The first, second, and third values are highlighted by orange, blue, and green backgrounds. All** 526
 469 **numbers are denoted in % with the maximum value of 100%.** 527

Model	Candle.	Tech.	Stat.	Tab.	Doc.	Mixed	CS	MP	VS	HS					
Open source MLLMs															
Yi-VL-34B [10]	23.64	16.36	18.76	15.42	14.89	32.38	19.42	14.39	26.06	16.62					
CogVLM2-19B [11]	39.44	35.57	52.30	50.38	45.76	57.14	47.33	44.22	49.70	49.09					
InternVL2-2B [12]	30.35	33.18	38.62	40.00	38.49	58.10	40.36	34.73	35.45	34.55					
InternVL2-4B [12]	35.38	38.98	51.48	54.66	47.77	63.81	50.87	44.85	43.64	45.71					
InternVL2-8B [12]	42.38	45.00	60.41	57.79	52.59	67.62	56.39	51.56	48.79	49.87					
InternVL2-76B [12]	55.52	47.50	63.02	70.84	63.09	67.62	62.78	61.73	54.54	58.70					
InternVL3-2B [7]	38.32	47.16	58.42	57.86	53.53	62.86	55.09	52.19	46.97	49.09					
InternVL3-8B [7]	53.29	60.91		70.65	71.30	64.75	63.81	68.88	63.80	56.97	61.82				
InternVL3-9B [7]	52.87	60.00	71.48	72.21	64.39	67.62	69.06	63.59	57.88	64.16					
InternVL3-14B [7]	59.44	61.59	73.06		77.25		67.27	60.95	70.76	67.89		62.12	69.35		
InternVL3-38B [7]	61.12	64.32	68.52	76.26	65.18	59.05	69.35	65.86	60.30	74.29					
InternVL3-78B [7]	64.90		64.32	70.65	78.32		73.53		61.90	72.96	69.45		63.33	76.62	
LLaMA3.2-11B [13]	35.24	31.59	47.63	50.92	39.42	48.57	45.16	39.07	38.79	47.79					
LLaMA3.2-90B [13]	40.56	40.11	51.20	58.17	45.83	64.76	50.14	46.33	46.06	56.10					
LLaVA-Next-7B [14]	29.65	23.52	28.80	28.32	28.34	44.76	28.45	26.08	32.73	35.32					
LLaVA-Next-13B [14]	27.27	26.36	33.68	32.14	32.95	39.05	32.67	29.20	30.91	35.84					
MiniCPM2.6 [15]	45.03	45.00	54.23	58.63	49.42	59.05	52.09	50.51	45.45	60.78					
Phi3-Vision [16]	37.62	40.00	49.48	49.54	48.71	62.86	49.75	43.08	40.30	52.21					
Phi3.5-Vision [16]	32.73	30.45	46.25	38.24	39.21	59.05	44.73	32.28	41.52	36.88					
Qwen2VL-2B [17]	38.74	40.80	46.60	46.26	44.68	57.14	45.13	43.71	38.79	48.57					
Qwen2VL-7B [17]	39.72	41.70	46.60	46.11	44.03	54.29	44.73	44.09	36.97	50.91					
Qwen2VL-72B [17]	60.12	60.11	65.15	71.73	66.04	74.24	67.65	62.78	68.48	67.01					
Qwen2.5VL-3B [8]	50.07	54.32	61.79	62.14	55.68	60.95	61.05	54.81	49.39	60.78					
Qwen2.5VL-7B [8]	55.24	57.61	64.95	66.64	60.79	61.90	62.53	61.14	60.30	64.94					
Qwen2.5VL-32B [8]	60.98	59.77	67.70	72.44	62.37	63.81	67.40	63.25	60.00	69.09					
Qwen2.5VL-72B [8]	62.38	58.86	69.76	77.02	67.48	63.81	70.54	66.08	57.88	73.25					
Proprietary MLLMs															
GeminiPro1.5 [18]	51.19	57.39	65.09	69.24	58.92	73.33	64.91	58.31	56.67	65.97					
Claude3.5-Sonnet [9]	51.47	53.52	72.51	71.15	59.93	79.05		67.47	58.19	69.70		68.57			
GPT-4o-mini [6]	58.18	55.80	70.38	66.56	64.03	76.00		66.44	60.89	63.94	72.21				
GPT-4o [6]	67.27		70.45	75.19	75.11	72.16		76.19	76.06	66.84	74.85		84.16		
Gemini2.5Pro [5]	72.73		71.82	82.61	85.80	78.13	74.29	81.08	76.92	74.24	85.19				

Reasoning

- (1) **Accurate Numerical Calculation:** Perform accurate numerical calculation or numerical comparison based on the numbers presented in the image.
- (2) **Estimated Numerical Calculation:** Obtain approximate values based on relevant clues(such as spatial location) and perform numerical calculation.

Cognition

- (1) **Risk Warning:** Give an investment risk based on the information in the image (and background information).

- (2) **Investment Advice:** Give an investment advice based on the information in the image (and the background information).
- (3) **Explain Reason:** Give an reason for the phenomenon indicated in the question based on the information in the image (and background information).
- (4) **Financial Question Answer:** Answer objective financial questions based on general financial knowledge.

Hallucination

Table 5: Evaluation results in the Chinese MME-Finance for all tasks. Abbreviations and color settings are the same as before.
The first, second, and third highest values are highlighted by orange, blue, and green backgrounds. All numbers are denoted in % with the maximum value of 100%.

Model	Overall	Perception			Reasoning			Cognition			NA
		IC	OCR	ER	SA	ANC	ENC	RW	IA	RE	
Open source MLLMs											
Yi-VL-34B [10]	23.50	43.89	0.66	9.86	4.94	23.97	18.13	20.00	28.79	60.00	51.81
CogVLM2-19B [11]	35.32	55.69	41.10	37.84	16.02	39.37	29.38	8.11	31.43	26.15	28.47
InternVL2-2B [12]	50.06	68.06	68.57	45.68	24.94	45.56	39.38	59.46	47.25	67.69	51.39
InternVL2-4B [12]	45.78	67.22	60.22	43.51	19.28	51.43	38.75	31.89	46.59	63.08	36.94
InternVL2-8B [12]	58.44	73.47	76.92	55.14	25.18	52.84	42.50	53.51	61.32	76.92	67.78
InternVL2-76B [12]	62.63	73.47	75.71	61.35	38.43	64.13	53.13	58.38	63.08	75.38	67.36
InternVL3-2B [7]	50.06	68.06	68.57	45.68	24.94	45.56	39.38	59.46	47.25	67.69	51.39
InternVL3-8B [7]	69.21	75.83	86.81	68.65	36.87	74.92	63.13	77.30	68.57	81.54	73.06
InternVL3-9B [7]	69.12	75.69	85.38	68.78	39.64	75.24	61.25	70.81	70.11	83.08	74.44
InternVL3-14B [7]	71.91	76.39	90.55	65.27	49.28	79.37	68.13	67.57	67.69	84.62	74.03
InternVL3-38B [7]	74.07	76.25	90.88	75.54	48.43	81.43	70.63	77.84	68.35	84.62	78.06
InternVL3-78B [7]	73.62	74.17	91.87	69.59	48.80	82.38	78.13	80.00	67.25	75.38	77.50
LLaVA-Next-7B [14]	21.45	50.69	8.35	12.16	9.28	16.03	13.13	12.43	28.35	46.15	25.14
LLaVA-Next-13B [14]	19.87	49.58	8.68	12.30	13.01	14.60	9.38	8.11	24.84	13.85	17.64
MiniCPM2.6 [15]	38.60	53.47	64.29	45.27	23.98	18.41	27.50	32.43	36.70	35.38	27.92
Phi3-Vision [16]	31.91	57.92	32.31	40.68	16.02	29.05	23.13	22.70	32.31	43.08	14.31
Phi3.5-Vision [16]	30.12	55.97	19.45	20.27	23.85	20.48	24.38	28.65	41.98	41.54	26.94
Qwen2VL-2B [17]	49.12	65.97	63.41	48.38	24.94	39.05	36.88	36.76	46.37	56.92	51.53
Qwen2VL-7B [17]	64.91	73.61	84.95	64.05	34.34	69.68	58.13	55.14	59.34	67.69	65.97
Qwen2VL-72B [17]	73.35	79.58	89.67	73.24	55.90	73.81	73.13	69.19	65.05	76.92	74.17
Qwen2.5VL-3B [8]	61.38	72.22	78.68	55.41	39.64	65.08	58.13	51.35	55.38	75.38	60.56
Qwen2.5VL-7B [8]	69.94	75.56	88.68	67.84	49.40	73.02	71.25	64.86	64.62	73.85	67.08
Qwen2.5VL-32B [8]	70.43	76.53	86.48	70.27	48.19	72.06	62.50	80.00	65.93	90.77	68.06
Qwen2.5VL-72B [8]	76.95	78.33	94.73	75.54	54.46	81.11	73.13	83.78	70.99	84.62	77.92
Proprietary MLLMs											
Claude3.5-Sonnet [9]	71.04	74.03	74.73	77.43	51.33	76.83	66.25	79.46	68.79	81.54	70.28
GeminiPro1.5 [18]	69.25	75.42	81.43	67.30	44.58	76.03	78.75	74.59	62.42	66.15	74.58
GPT-4o-mini [6]	54.58	66.25	59.01	47.84	30.60	50.63	54.38	63.78	58.68	72.31	67.36
GPT-4o [6]	59.53	69.86	68.90	58.92	39.04	57.46	55.63	60.54	53.19	63.08	68.19
Gemini2.5Pro [5]	81.07	76.11	96.04	85.41	67.23	86.98	87.88	83.78	72.09	87.69	77.08

- (1) **Not Applicable:** The answer is not available based on the content of the image or the general world knowledge stored in MLLMs.

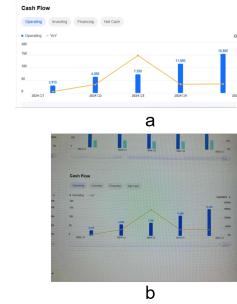


Figure 5: Comparison of the difficulty of recognizing computer screenshot versus photos taken with a mobile phone.

Q. How much did the cash flow increase from 2024 Q3 to 2025 Q1?

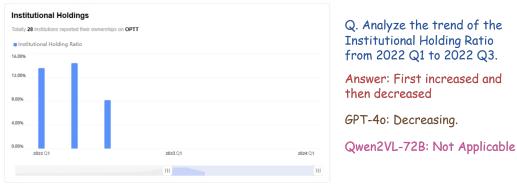
Answer: The cash flow in 2024 Q3 is 7.33B, and the cash flow in 2025 Q1 is 15.35B, hence the increase is 15.35-7.33=8.02B.

Prediction(a): The cash flow for 2024 Q3 is 7.33B, and for 2025 Q1, it is 15.35B. The calculation is as follows: 15.35B - 7.33B = 8.02B.

Prediction(b): The cash flow for 2024 Q3 is 7.33B and for 2025 Q1 is 15.35B. The calculation is as follows: 15.35B - 7.33B = 8.02B.

Table 6: Evaluation results on Chinese MME-Finance for different types and styles of images. Abbreviations and color settings are the same as before. The first, second, and third highest values are highlighted by orange, blue, and green backgrounds. All numbers are denoted in % with the maximum value of 100%.

Model	Candle.	Tech.	Stat.	Tab.	Doc.	Mixed	CS	MP	VS	HS
Open source MLLMs										
Yi-VL-34B [10]	26.00	21.66	22.87	23.55	23.63	24.52	26.30	20.49	25.09	22.22
CogVLM2-19B [11]	38.86	33.59	31.19	37.74	37.48	37.48	37.11	32.50	36.77	39.17
InternVL2-2B [12]	36.29	35.80	39.01	37.90	48.22	32.26	43.70	36.92	39.63	34.72
InternVL2-4B [12]	34.43	44.20	46.93	48.31	51.56	37.10	51.90	40.80	43.72	45.56
InternVL2-8B [12]	49.71	55.69	56.44	59.60	66.44	53.23	61.80	55.31	62.11	50.00
InternVL2-76B [12]	55.86	64.75	61.39	62.74	67.56	53.87	65.55	57.28	69.44	63.61
InternVL3-2B [7]	35.86	53.15	49.80	47.02	58.81	48.06	52.04	47.77	52.92	46.39
InternVL3-8B [7]	63.43	67.40	73.07	68.15	72.59	64.52	72.04	65.27	72.80	69.17
InternVL3-9B [7]	60.29	65.08	70.50	73.71	72.59	62.90	73.22	66.03	69.81	62.78
InternVL3-14B [7]	60.14	69.94	74.16	76.94	75.41	61.61	75.12	68.08	73.66	73.06
InternVL3-38B [7]	61.43	73.48	76.73	79.11	76.07	66.77	77.49	71.12	75.16	70.00
InternVL3-78B [7]	64.14	72.38	77.92	76.61	74.89	67.10	77.30	70.00	73.42	75.00
LLaVA-Next-7B [14]	29.57	20.88	20.00	16.21	25.33	13.55	22.89	19.42	23.23	21.67
LLaVA-Next-13B [14]	24.14	21.66	21.39	16.37	19.56	15.48	20.33	19.29	19.88	20.83
MiniCPM2.6 [15]	36.57	36.69	37.92	40.08	44.81	18.06	38.58	35.80	47.33	36.67
Phi3-Vision [16]	31.71	35.47	29.21	31.45	34.30	22.26	36.45	27.32	31.55	34.72
Phi3.5-Vision [16]	30.29	35.03	27.92	25.48	33.33	27.10	32.09	27.54	29.81	35.28
Qwen2VL-2B [17]	40.71	42.10	49.11	49.68	59.41	41.61	49.95	48.88	51.80	39.72
Qwen2VL-7B [17]	55.71	60.55	69.21	66.13	68.37	64.52	69.29	62.41	62.73	59.72
Qwen2VL-72B [17]	64.14	71.71	77.52	75.65	75.26	67.74	76.35	69.96	74.53	74.17
Qwen2.5VL-3B [8]	44.14	59.34	63.37	65.16	67.78	56.77	64.08	61.61	56.40	55.28
Qwen2.5VL-7B [8]	60.57	65.86	72.38	72.50	74.15	66.45	73.51	69.42	66.83	59.17
Qwen2.5VL-32B [8]	64.00	72.82	68.32	68.15	77.48	63.23	73.70	67.37	70.56	70.00
Qwen2.5VL-72B [8]	60.00	77.90	79.90	81.21	80.96	68.39	80.95	74.11	78.14	68.61
Proprietary MLLMs										
Claude3.5-Sonnet [9]	68.71	71.27	72.28	71.13	72.22	66.13	73.22	64.46	80.12	78.89
GeminiPro1.5 [18]	66.43	73.04	73.76	63.15	71.78	63.23	72.23	64.51	74.16	70.28
GPT-4o-mini [6]	55.71	67.96	58.22	43.55	56.89	35.16	55.40	45.63	70.68	69.44
GPT-4o [6]	53.14	71.93	62.97	52.98	62.59	39.35	64.03	47.01	75.90	74.44
Gemini2.5Pro [5]	73.48	82.87	84.65	83.31	80.74	73.87	82.93	78.48	83.48	80.83



Q. Analyze the trend of the Institutional Holding Ratio from 2022 Q1 to 2022 Q3.
Answer: First increased and then decreased
GPT-4o: Decreasing.
Qwen2VL-72B: Not Applicable

Figure 6: The display of hallucination problems of MLLMs.

Table 7: Evaluation results in English MME-Finance with the prompt to allow “Not Applicable” response across all types of tasks. Abbreviations and color settings are the same as before. The first, second, and third highest values are highlighted by orange, blue, and green backgrounds. All numbers are denoted in % with the maximum value of 100%.

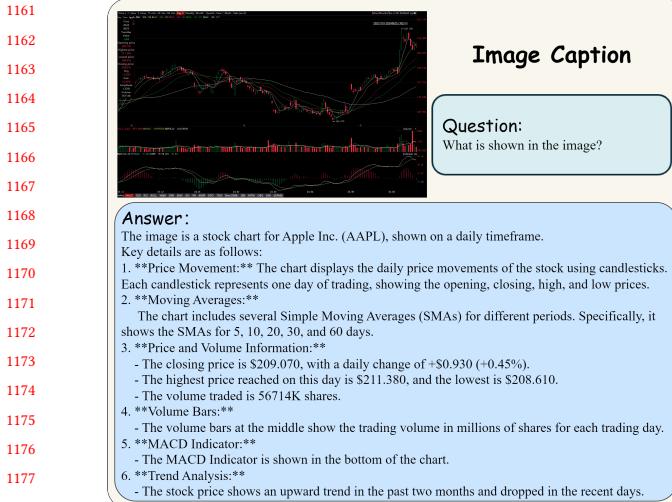
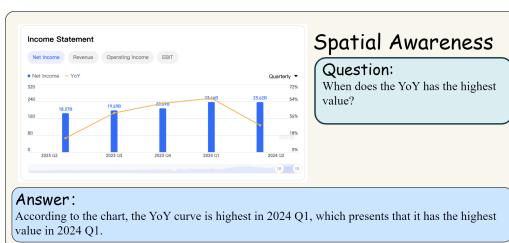
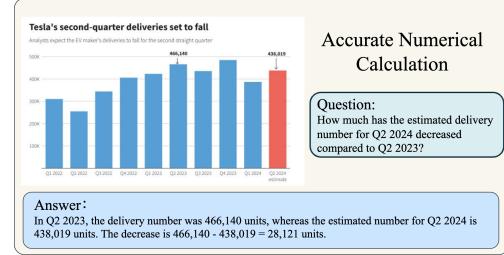
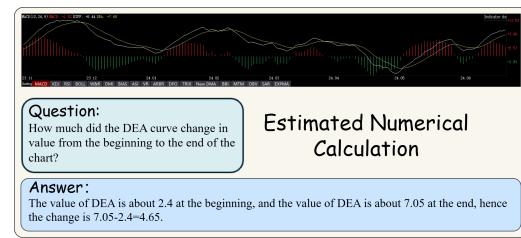
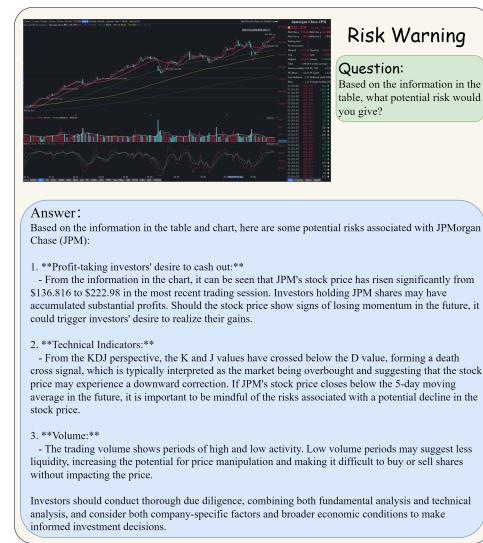
Model	Overall	Perception			Reasoning			Cognition			NA
		IC	OCR	ER	SA	ANC	ENC	RW	IA	RE	
Open source MLLMs											
CogVLM2-19B [11]	31.24	36.22	42.02	26.99	7.95	34.14	19.52	13.64	38.11	32.22	48.44
InternVL2-2B [12]	32.16	61.22	32.81	12.76	4.72	32.78	21.43	54.55	47.92	63.33	50.20
InternVL2-4B [12]	45.93	68.05	54.27	28.22	18.69	54.74	32.38	68.18	50.19	68.89	59.46
InternVL2-8B [12]	50.59	70.00	60.11	33.99	23.84	62.56	40.00	76.36	60.75	75.56	58.37
LLaVA-Next-7B [14]	20.10	58.05	6.18	2.70	1.31	6.62	3.33	43.64	37.74	16.67	43.27
MiniCPM2.6 [15]	48.37	69.63	62.81	34.85	21.31	54.89	28.57	50.00	40.38	45.56	62.31
Phi3-Vision [16]	37.06	69.51	58.43	29.57	11.88	22.11	3.81	7.27	16.60	41.11	47.76
Phi3.5-Vision [16]	28.69	66.83	33.03	13.87	8.12	17.14	6.19	2.73	15.47	22.22	45.58
Qwen2VL-2B [17]	32.47	62.32	40.45	9.08	6.72	41.80	25.24	31.82	36.23	13.33	40.14
Qwen2VL-7B [17]	32.40	62.32	40.00	8.10	7.07	41.80	22.86	33.64	40.00	10.00	40.14
Qwen2VL-72B [17]	62.97	80.49	83.26	50.43	25.50	78.95	46.67	73.64	68.30	73.33	73.06
											86.36

Table 8: Evaluation results on English MME-Finance for different types and formats of images with the prompt to allow “Not Applicable” response across all types of tasks. Abbreviations and color settings are the same as before. The first, second, and third highest values are highlighted by orange, blue, and green backgrounds. All numbers are denoted in % with the maximum value of 100%.

Model	Candle.	Tech.	Stat.	Tab.	Doc.	Mixed	CS				HS
							MP	VS	MP	VS	
Open source MLLMs											
CogVLM2-19B [11]	25.31	19.89	33.13	32.52	37.53	39.05	31.70	29.66	33.33	35.84	
InternVL2-2B [12]	25.03	25.57	36.64	32.37	33.88	53.33	36.21	28.82	25.76	29.09	
InternVL2-4B [12]	30.91	34.43	54.35	45.47	52.37	56.19	49.68	43.63	39.39	38.70	
InternVL2-8B [12]	39.02	38.18	55.27	50.72	58.08	69.52	54.98	47.13	49.70	41.04	
LLaVA-Next-7B [14]	16.08	18.75	22.82	19.31	19.93	33.33	22.42	16.71	22.12	22.60	
MiniCPM2.6 [15]	39.86	44.66	51.53	48.85	52.23	38.10	47.26	48.48	48.48	55.58	
Phi3-Vision [16]	24.62	28.18	42.82	40.00	42.16	12.38	38.66	35.86	29.70	39.22	
Phi3.5-Vision [16]	20.98	22.95	30.23	30.29	34.36	10.48	32.92	22.95	31.82	30.91	
Qwen2VL-2B [17]	22.80	24.32	35.34	37.63	33.20	48.57	33.86	30.42	29.39	36.62	
Qwen2VL-7B [17]	23.92	23.86	35.57	37.91	33.06	43.81	33.54	31.18	29.39	35.32	
Qwen2VL-72B [17]	52.31	56.14	69.16	64.32	64.67	74.29	65.78	59.41	63.94	63.90	

929	STEP1: We provide a general prompt: 'You are a financial expert who is well - versed in various financial	987
930	charts and has extensive financial knowledge. Now you are given an image and a corresponding question.	988
931	Please answer this question.'	989
932		990
933	STEP2: We present the question: 'This is question: + content of the question'	991
934		992
935	STEP3: We state the requirements for answering the question. Here, for different types of questions, the	993
936	answering requirements are also different: 'Here are the answer requirements: + answering requirements'	994
937	The answering requirements for various questions are as follows:	995
938	' Image Caption ': 'Describe this image in a whole - part structure. Start with a sentence summarizing the	996
939	main theme of the image. If the image depicts multiple objects, first introduce each object in one	997
940	sentence, and if there are some connections between objects, explain each connection in one sentence. If	998
941	the object is complex, it can be further explained. Your answer should be less than 250 words and should	999
942	not include any irrelevant information.'	1000
943	' OCR ': 'The answer should conduct an Optical Character Recognition (OCR) analysis on the content inquired	1001
944	about. Just answer the question with a single word or phrase if possible. No irrelevant information should	1002
945	be included.'	1003
946	' Entity Recognition ': 'The answer should contain recognition results of entities mentioned in the question.	1004
947	Just answer the question with a single word or phrase if possible. No irrelevant information should be	1005
948	included.'	1006
949	' Spatial Awareness ': 'The answer should be based on the spatial relationships between entities in the	1007
950	question. It is best to provide corresponding evidence for all judgments. If specific numerical answers are	1008
951	not present in the image but can be estimated based on its content, the estimated results can be used. Just	1009
952	provide the answer in one word or a short sentence. No irrelevant information should be included.'	1010
953	' Numerical Calculation ': 'You should perform mathematical calculations based on the information in the	1011
954	image. You need to estimate some values that are not directly displayed in the image for answering the	1012
955	question. You should show the calculation process and output the calculated result.'	1013
956	' Accurate Numerical Calculation ': 'You should perform mathematical calculations based on the information	1014
957	in the image. You need to provide a step - by - step calculation and obtain a numerical result.'	1015
958	' Financial Knowledge ': 'The answer should be based on financial knowledge. Briefly answer the question	1016
959	within 100 words. The answer should not contain irrelevant content related to the picture.'	1017
960	' Risk Warning ': 'You should warn of investment risk based on the information in the chart and professional	1018
961	financial knowledge. All your arguments need to be supported by facts or theories and the answer should be	1019
962	within 150 words.'	1020
963	' Investment Advice ': 'You should provide investment advice based on the information in the chart and	1021
964	professional financial knowledge. All your arguments need to be supported by facts or theories and the	1022
965	answer should be within 150 words.'	1023
966	' Explain Reason ': 'You should provide an explanation based on the information in the chart and professional	1024
967	financial knowledge. All your arguments need to be supported by facts or theories and the answer should be	1025
968	within 150 words.'	1026
969	' Not Applicable ': 'If you cannot answer, please say "Not Applicable", and provide the explanations.'	1027
970	STEP4: If the question has background information, we will add a background description for it: 'This is	1028
971	background: + background description'	1029
972		1030
973		1031
974	Figure 7: Inference prompt.	1032
975		1033
976		1034
977		1035
978		1036
979		1037
980		1038
981		1039
982		1040
983		1041
984		1042
985		1043
986		1044

1045		
1046	STEP1: We provide a general prompt: "You are a financial expert who is well - versed in various financial charts and has extensive financial knowledge. Now you are given an image and a corresponding question. Please answer this question."	1103
1047		1104
1048		1105
1049	STEP2: Then we present the scoring rules. Depending on the type of question, questions can be divided into objective questions, accurate calculation questions, numerical calculation questions, subjective questions, and not Applicable question, and they each have their own prompts, as shown below:	1106
1050	" objective question ": "This is an objective question. Please give a score: (The full score is 5 points in total. Score according to the following conditions.)	1107
1051	Answer accuracy: Full score is 5 points. In combination with the question, it is required that the content and semantics of the prediction and the answer must be the same and there should be no redundant answers. The answer can be expressed in different ways, such as different unit symbols and different counting methods. If the answer is correct, 5 points can be given. If the answer contains multiple pieces of content, multiply 5 by the correct proportion of the prediction to give the final score. If the answer is wrong, give 0 points directly."	1108
1052	" accurate calculation question ": "This is a calculation question. Please give a score: (The full score is 5 points in total. Score item by item according to the following conditions and add up the obtained scores to get the total score.)	1109
1053	1. Answer accuracy: Full score is 2 points. In combination with the question and answer, it is required that the final calculated result of the prediction must be accurate. If the answer is correct, give 2 points. If the answer is wrong, give 0 points.	1110
1054	2. Calculation process: Full score is 3 points. There should be intermediate calculation processes for calculation questions, and they should also be correct. In combination with the answer, if all elements and steps are included in the prediction, give 3 points. If the final answer is wrong but the calculation process included in the prediction is partially correct, multiply 3 by the correct proportion to give the final score. If the calculation process is also wrong, give 0 points."	1111
1055	" numerical calculation question ": "This is a valuation calculation question. Please give a score: (The full score is 5 points in total. Score item by item according to the following conditions and add up the obtained scores to get the total score.)	1112
1056	1. Answer accuracy: Full score is 2 points. In combination with the question and answer, if the predicted final result fluctuates within $\pm 10\%$ of the final value of the answer, give 2 points. If the predicted final result fluctuates within $\pm 10\% - \pm 20\%$ of the final value of the answer, give 1 point. If the predicted final result fluctuates more than $\pm 20\%$ of the final value of the answer, give 0 points.	1113
1057	2. Calculation process: Full score is 3 points. There should be intermediate calculation processes for calculation questions, and they should also be correct. In combination with the answer, if all elements and steps are included in the prediction, give 3 points. If the final answer is wrong but the calculation process included in the prediction is partially correct, multiply 3 by the correct proportion to give the final score. If the calculation process is also wrong, give 0 points."	1114
1058	" subjective question ": "This is a subjective question. Please give a score: (The full score is 5 points in total. Score item by item according to the following conditions and add up the obtained scores to output the total score.)	1115
1059	1. Content matching degree: Full score is 2 points. When all keywords of the answer appear in the predicted text, give 2 points. When some keywords of the answer appear in the predicted text, give 1 point. When none of the keywords of the answer appear in the predicted text, give 0 points.	1116
1060	2. Semantic matching degree: Full score is 2 points. When the semantics of the answer and the predicted content are close and there is no wrong judgment, give 2 points. When part of the semantics of the answer and the predicted content are close, give 1 point. When the semantics of the answer and the predicted content are completely different, give 0 points.	1117
1061	3. Problem attribute self-consistency: Full score is 1 point. Give points as appropriate according to the following prediction requirements, and require smooth logic and correct grammar."	1118
1062	" not Applicable question ": "This is an unanswerable question. Please give a score: (The full score is 5 points in total. Score according to the following conditions.)	1119
1063	Answer accuracy: Full score is 5 points. If 'Not Applicable' appears in the prediction result, give 5 points directly. If 'Not Applicable' does not appear in the prediction but indicates that it cannot be answered, give points as appropriate."	1120
1064		1121
1065		1122
1066		1123
1067		1124
1068		1125
1069		1126
1070		1127
1071		1128
1072		1129
1073		1130
1074		1131
1075		1132
1076		1133
1077		1134
1078		1135
1079		1136
1080		1137
1081		1138
1082		1139
1083		1140
1084		1141
1085		1142
1086		1143
1087		1144
1088	At the same time, we will also add the response requirements during inference as prompts (refer to the inference response requirements).	1145
1089		1146
1090	STEP3: For different types of questions, we will also give some scoring examples. This can provide some few - shot for the scoring model. The following is an example of 'OCR':	1147
1091	"Just complete the last space of the correctness score.	1148
1092	Question Ground truth Prediction Correctness	1149
1093	--- --- --- ---	1150
1094	What is the operating cash flow value in 2024 Q4? 11.50 billion 11.50B 5	1151
1095	What are the last candlestick values for SMA 5/10/20/30/60? The last candlestick values for the Simple Moving Averages (SMA) of periods 5, 10, 20, 30, and 60 for Microsoft Corp. (MSFT) are as follows: - SMA 5: 448.09 - SMA 10: 444.67 - SMA 20: 433.29 - SMA 30: 430.03 - SMA 60: 420.50 5:448.09, 10:444.67, 30:430.03, 60:420.50 4 "	1152
1096		1153
1097		1154
1098	STEP4: Finally, we will arrange the questions, answers, and predictions in the same format as the above examples for the scoring model to score. The following is an example of 'OCR' question:	1155
1099	"What is the Turnover shown in the chart? 3.48B 3.48B "	1156
1100		1157
1101		1158
1102		1159

**Figure 9: Image Caption.****Figure 10: OCR.****Figure 11: Entity Recognition.****Figure 12: Spatial Awareness.****Figure 13: Accurate Numerical Calculation.****Figure 14: Estimated Numerical Calculation.****Figure 15: Risk Warning.**

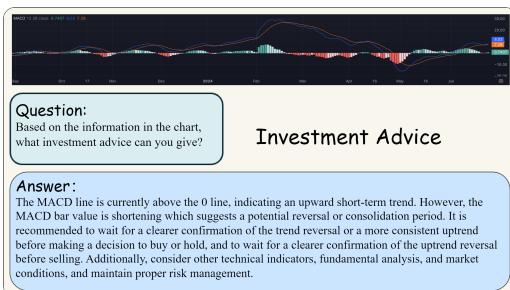


Figure 16: Investment Advice.

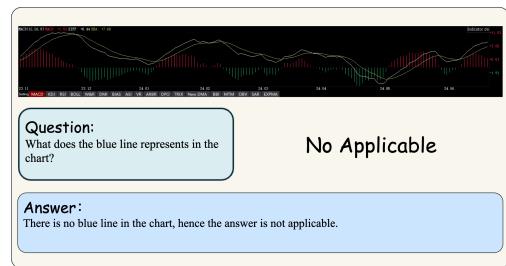


Figure 19: Not Applicable.

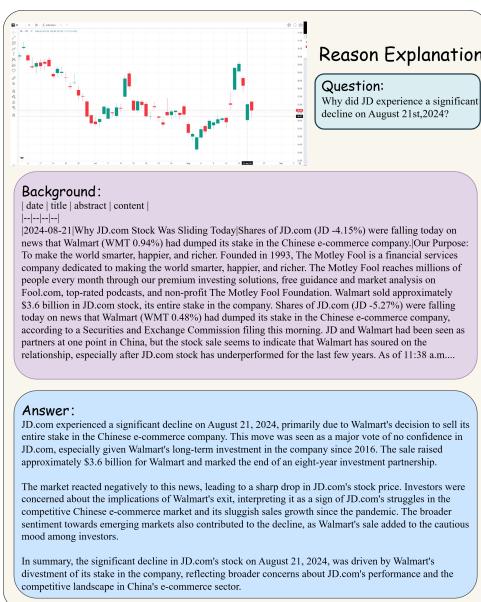


Figure 17: Reason Explanation.

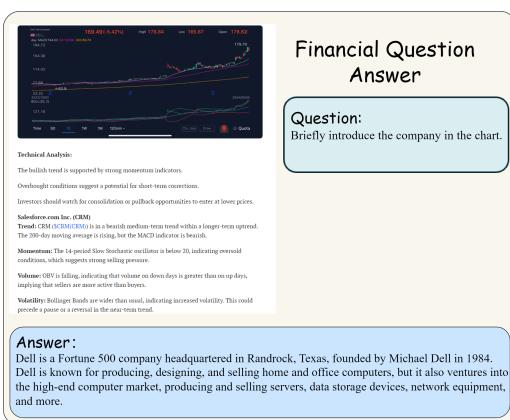


Figure 18: Financial Question Answer.

1393 2 Binary-choice Version

1394 We design binary-choice questions to objectively evaluate the model's
 1395 perception capabilities with respect to financial images. In this ver-
 1396 sion, all questions are Chinese.

1397

1398 2.1 Data Construction Pipeline

1399
1400
1401
1402
1403
1404
1405
1406
1407

First, we collected a set of financial images, some of which contained circled markings. Subsequently, we utilize GPT-4o [6] to generate binary-choice questions related to these circled marks based on the images (e.g. Does the image contain circled marks? Is the xx content within the circled area in the image?). GPT-4o also provided preliminary answers to these questions. Subsequently, our annotators reviewed the generated questions and answers, eliminating any unreasonable questions and revising any incorrect answers.

1408
1409

2.2 Experimental Result

1410
1411
1412
1413
1414
1415

We report the accuracy and f1 score metrics of the models in Table 9. Since binary-choice questions and their examined content tend to focus on a lower-level understanding, it is evident that most models achieve relatively satisfactory results. Moreover, GPT-4o and Qwen2.5VL-72B demonstrate superior performance compared to other models in the two metrics.

1416
1417
1418

Table 9: Evaluation results on binary-choice version of MME-Finance.

1419

Model	Accuracy	F1-Score
Open source MLLMs		
InternVL3-2B [7]	68.45	77.88
InternVL3-8B [7]	75.00	81.05
InternVL3-9B [7]	68.95	78.43
InternVL3-14B [7]	71.10	79.92
InternVL3-38B [7]	76.90	83.31
InternVL3-78B [7]	81.30	86.26
Qwen2.5VL-3B [8]	69.85	79.33
Qwen2.5VL-7B [8]	83.90	87.14
Qwen2.5VL-32B [8]	78.60	84.47
Qwen2.5VL-72B [8]	85.40	88.73
Claude3.5-Sonnet [9]	82.40	86.94
GPT-4o-mini [6]	69.10	72.26
GPT-4o [6]	87.95	90.31
Gemini2.5Pro [5]	83.50	87.56

1437

1438

2.3 Example

1441
1442
1443
1444

As shown in Figure 20, there is a sub-image with a circled selection mark on the left, and the questions and answers are on the right. These questions evaluate the model's ability to understand the circled marks and the circled contents.

1445

1446

1447

1448

1449

1450

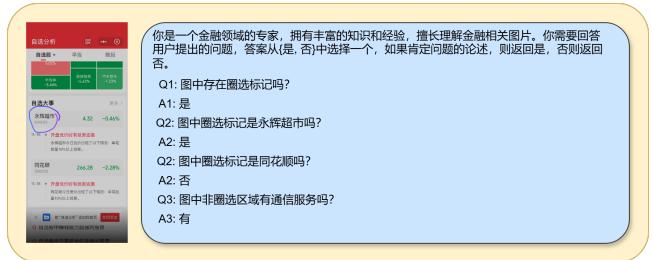


Figure 20: Example of the binary-choice version of MME-Finance.

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509 3 Multi-turn Version

1510 We design the multi-turn version to evaluate the model’s ability
 1511 in multi-turn multi-image financial scenarios. In this version, all
 1512 questions are Chinese.

1513 3.1 Data Construction Pipeline

1514 To ensure that the images are interconnected, we implemented the
 1515 following data collection strategy. First, a company was selected and
 1516 then its relevant financial images were captured, including K-line
 1517 charts, intraday charts, company profiles, industry reports, related
 1518 news articles, etc. Then, following the MMDU [19], we had GPT-
 1519 4o [6] generate multiple questions based on these images, ensuring
 1520 that most of the questions were associated with multiple images.
 1521 These questions will be reviewed by the annotators, and those
 1522 deemed inappropriate will be removed or revised. Subsequently,
 1523 the questions and images will be feed into GPT-4o to generate
 1524 preliminary responses, which will then undergo further refinement
 1525 to produce the final standard answers. Ultimately, we collected 48
 1526 sets of financial images from companies, each company providing
 1527 7 images. Each company was associated with 9 to 10 questions,
 1528 resulting in a total of 477 questions.

1529 3.2 Experimental Result

1530 We refer to MMDU [19] and evaluate the model’s responses from 6
 1531 dimensions, namely Creativity, Richness, Visual Perception, Logical
 1532 Coherence, Answer Accuracy, and Image Relationship Understanding.
 1533 In addition, the average score is adopted as a comprehensive
 1534 evaluation. As illustrated in Table 10, Gemini2.5Pro [5] achieves
 1535 the highest average score. Claude3.5-Sonnet and Qwen2.5VL-72B
 1536 achieve comparable performance. In terms of specific evaluation
 1537 dimensions, Gemini2.5Pro excels in creativity and richness, while
 1538 Qwen2.5VL-72B is superior in visual perception, logical consistency,
 1539 answer accuracy, and image relationship understanding.

1540 3.3 Example

1541 As shown in Figure 21, there are seven related financial subgraphs of
 1542 a company. Below are some multi-turn questions and their answers.
 1543 These questions require the model to distinguish each picture, col-
 1544 lect information from multiple pictures, and make reasonable infer-
 1545 ences to obtain the final answer comprehensively. These questions
 1546 serve as an effective evaluation of the model’s capability to integrate
 1547 information from multiple images and handle multi-turn questions
 1548 within the financial scenario.

Table 10: Evaluation results on multi-turn version of MME-Finance. We report the metrics of Creativity (C), Richness (R), Visual Perception (VP), Logical Coherence (LC), Answer Accuracy (AA), Image Relationship Understanding (IRU), and the averaged (Avg.) results.

Model	C	R	VP	LC	AA	IRU	Avg.
Open source MLLMs							
InternVL3-2B [7]	59.66	69.90	68.70	79.90	76.46	70.46	70.84
InternVL3-8B [7]	62.26	70.12	71.43	83.63	79.94	73.99	73.56
InternVL3-9B [7]	69.31	77.65	77.06	87.17	85.53	79.52	79.37
InternVL3-14B [7]	54.35	62.77	65.46	79.29	75.46	67.00	67.39
InternVL3-38B [7]	62.45	70.82	71.36	83.88	81.34	73.82	73.94
InternVL3-78B [7]	71.01	77.78	77.40	87.51	85.85	79.73	79.88
Qwen2.5VL-3B [8]	45.41	56.74	59.47	72.34	68.02	60.53	60.42
Qwen2.5VL-7B [8]	72.83	81.07	80.40	89.25	88.49	83.29	82.55
Qwen2.5VL-32B [8]	76.53	79.94	75.44	82.50	81.76	78.21	79.07
Qwen2.5VL-72B [8]	75.81	82.96	81.53	90.25	89.58	84.26	84.06
Claude3.5-Sonnet [9]	82.10	86.37	78.99	90.21	89.08	81.80	84.76
GPT-4o-mini [6]	59.01	63.72	45.44	73.66	63.28	46.74	58.64
GPT-4o [6]	63.50	69.79	58.83	80.23	72.85	61.74	67.82
Gemini2.5Pro [5]	84.59	87.15	80.69	90.08	87.69	84.07	85.71

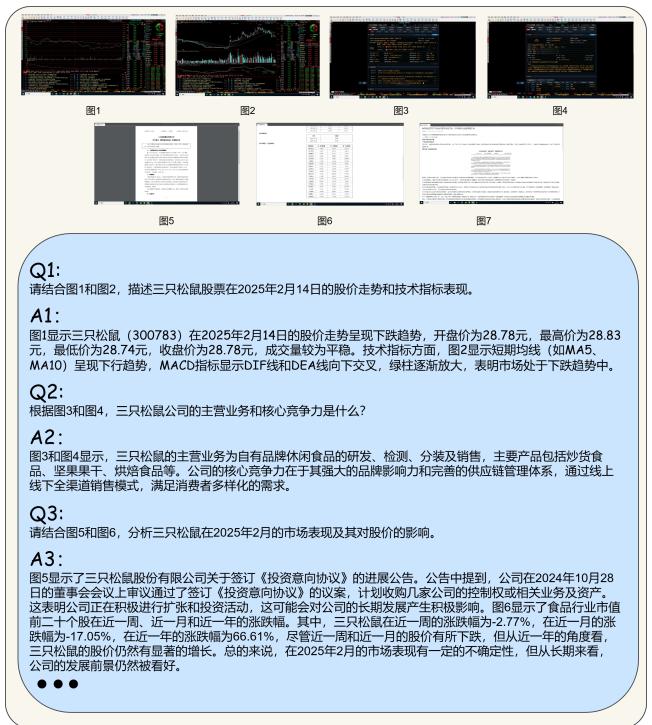


Figure 21: Example of the multi-turn version of MME-Finance.

References

- 1625
- 1626 [1] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque.
Chartqa: A benchmark for question answering about charts with visual and
logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- 1627 [2] Rengiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou,
Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile
benchmark and foundation model for complicated chart reasoning. *arXiv preprint
arXiv:2402.12185*, 2024.
- 1628 [3] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty.
Unichart: A universal vision-language pretrained model for chart comprehension
and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.
- 1629 [4] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa:
Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference
on Applications of Computer Vision*, pages 1527–1536, 2020.
- 1630 [5] Google. Gemini2.5pro, <https://deepmind.google/technologies/gemini/pro/>, 2025.
- 1631 [6] Open-AI. Gpt-4o, 2024.
- 1632 [7] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu,
Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced
training and test-time recipes for open-source multimodal models. *arXiv preprint
arXiv:2504.10479*, 2025.
- 1633 [8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai
Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report.
arXiv preprint arXiv:2502.13923, 2025.
- 1634 [9] Claude. Claude 3.5 sonnet, 2024.
- 1635 [10] 01. AI, ;, Alex Young, Bei Chen, Chao Li, Chenger Huang, Ge Zhang, Guanwei
Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng
Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao
Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong
Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi:
Open foundation models by 01.ai, 2024.
- 1636 [11] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang,
Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language
models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.
- 1637 [12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui,
Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to
gpt-4v? closing the gap to commercial multimodal models with open-source
suites. *arXiv preprint arXiv:2404.16821*, 2024.
- 1638 [13] Meta. Llama3.2, 2024.
- 1639 [14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and
Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January
2024.
- 1640 [15] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi
Cai, Haoyu Li, Weilin Zhao, Zihui He, et al. Minicpm-v: A gpt-4v level mllm
on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- 1641 [16] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed
Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harki-
rat Behl, et al. Phi-3 technical report: A highly capable language model locally
on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- 1642 [17] Qwen team. Qwen2-vl, 2024.
- 1643 [18] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol
Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini
1.5: Unlocking multimodal understanding across millions of tokens of context.
arXiv preprint arXiv:2403.05530, 2024.
- 1644 [19] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian
Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdm: A multi-turn multi-
image dialog understanding benchmark and instruction-tuning dataset for lvms.
arXiv preprint arXiv:2406.11833, 2024.
- 1645
- 1646
- 1647
- 1648
- 1649
- 1650
- 1651
- 1652
- 1653
- 1654
- 1655
- 1656
- 1657
- 1658
- 1659
- 1660
- 1661
- 1662
- 1663
- 1664
- 1665
- 1666
- 1667
- 1668
- 1669
- 1670
- 1671
- 1672
- 1673
- 1674
- 1675
- 1676
- 1677
- 1678
- 1679
- 1680
- 1681
- 1682
- 1683
- 1684
- 1685
- 1686
- 1687
- 1688
- 1689
- 1690
- 1691
- 1692
- 1693
- 1694
- 1695
- 1696
- 1697
- 1698
- 1699
- 1700
- 1701
- 1702
- 1703
- 1704
- 1705
- 1706
- 1707
- 1708
- 1709
- 1710
- 1711
- 1712
- 1713
- 1714
- 1715
- 1716
- 1717
- 1718
- 1719
- 1720
- 1721
- 1722
- 1723
- 1724
- 1725
- 1726
- 1727
- 1728
- 1729
- 1730
- 1731
- 1732
- 1733
- 1734
- 1735
- 1736
- 1737
- 1738
- 1739
- 1740