

---

## SUPPLEMENTARY MATERIAL OVERVIEW

- Section **A**: Related work.
- Section **B**: Automated data curation pipeline.
- Section **C**: Additional dataset details.
- Section **D**: Future direction.
- Section **E**: Additional experiments and analysis.
- Section **F**: Additional experimental details.
- Section **G**: More dataset details.
- Section **H**: Qualitative examples.

## A RELATED WORK

**Large Multimodal Models.** Recently, multimodal models (Radford et al., 2021; Li et al., 2022; OpenAI, 2023; Rombach et al., 2022; Jiang et al., 2024b; Zong et al., 2024a) has gained unparalleled attention. Building on the success of Large Language Models (LLMs) (Touvron et al., 2023a;b) and large-scale vision models (Radford et al., 2021), Large Multimodal Models (LMMs) are gaining prominence across diverse domains. These models extend LLMs to handle tasks involving various modalities, including mainstream 2D image processing (Liu et al., 2023a; Zhu et al., 2023; Lin et al., 2023; Gao et al., 2023; Ma et al., 2025), as well as 3D point clouds (Xu et al., 2023; Guo et al., 2023; 2024; Tang et al., 2025; Guo et al., 2025a), medical images (Xia et al., 2024c;b), videos (Li et al., 2023; Chen et al., 2023a; Zhang et al., 2023; Fu et al., 2024), reasoning (Guo et al., 2025b; Jiang et al., 2025), and robotics (Liu et al., 2024b; Jia et al., 2024). Among these LMMs, OpenAI’s GPT-4o (OpenAI, 2024b) and Anthropic’s Claude 3.5 Sonnet (Anthropic, 2024) demonstrate outstanding visual reasoning and comprehension capability, setting new standards in multi-modal performance. However, their closed-source nature limits broader adoption and development. In contrast, another research trajectory focuses on open-source LMMs for the community. Pioneering works like LLaVA (Liu et al., 2023a; 2024a; Li et al., 2024b;a), LLaMA-Adapter (Zhang et al., 2024b; Gao et al., 2023), and MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023b) incorporate a frozen CLIP (Radford et al., 2021) model for image encoding and integrate visual information into LLM for multi-modal instruction tuning. Later, works such as mPLUG-Owl (Ye et al., 2023a;b; 2024), SPHINX (Gao et al., 2024; Lin et al., 2023), and InternLM-XComposer (Dong et al., 2024) further advanced the field by incorporating diverse visual instruction tuning data and generalizing to more scenarios. More recent developments in the field have taken diverse directions. For example, several studies (Zong et al., 2024b; Tong et al., 2024) explore multiple vision encoders design. Meanwhile, other works (Liu et al., 2024a; Chen et al., 2024c; Qwen Team, 2024) incorporate high-resolution image input. Multi-image instruction data (Li et al., 2024b; Jiang et al., 2024a) is also integrated to enable perception across multiple images. While various benchmarks, both in the general (Fu et al., 2023; Jiang et al., 2025; Liu et al., 2023b; Yu et al., 2023; Xia et al., 2024a; Hong et al., 2025) and expert (Zhang et al., 2024c; Lu et al., 2023; 2022; Xia et al.) domain, has been proposed, the potential of LMM to function as a multimodal search engine remains largely unexplored. To this end, we introduce the MMSEARCH benchmark, which evaluates LMMs’ zero-shot abilities of multimodal search, offering valuable insights for future research.

**Large models with Retrieval Augmented Generation (RAG).** RAG (Retrieval-Augmented Generation) is an effective strategy for enhancing model knowledge by retrieving relevant information from external sources (Fan et al., 2024). RAG has been leveraged in various scenarios including knowledge-intensive question answering (Borgeaud et al., 2022; Guu et al., 2020), machine translation (He et al., 2021), and hallucination elimination (Bécharde & Ayala, 2024). Current works has focused on improving specific aspects of RAG. RG-RAG (Chan et al., 2024) proposes to refine the query for retrieval by decomposition and disambiguation. Self-RAG (Asai et al., 2023) incorporates the self-reflection of LLM to enhance the generation quality. The AI search engine could be viewed as a form of RAG with the Internet serving as the external knowledge source. Recently, MindSearch (Chen et al., 2024b) proposes an AI search engine framework to simulate the human minds in web information seeking. Meanwhile, multiple benchmarks of RAG (Yang et al., 2024;

Chen et al., 2024a) have been introduced to comprehensively evaluate a RAG system. However, both the current AI search engine and RAG benchmark are limited to the text-only setting, leaving the multimodal search engine and evaluation largely unexplored. To bridge this gap, we introduce MMSEARCH-ENGINE and MMSEARCH, a multimodal AI search engine pipeline and dataset designed to evaluate various multimodal scenarios.

## B AUTOMATED DATA CURATION PIPELINE

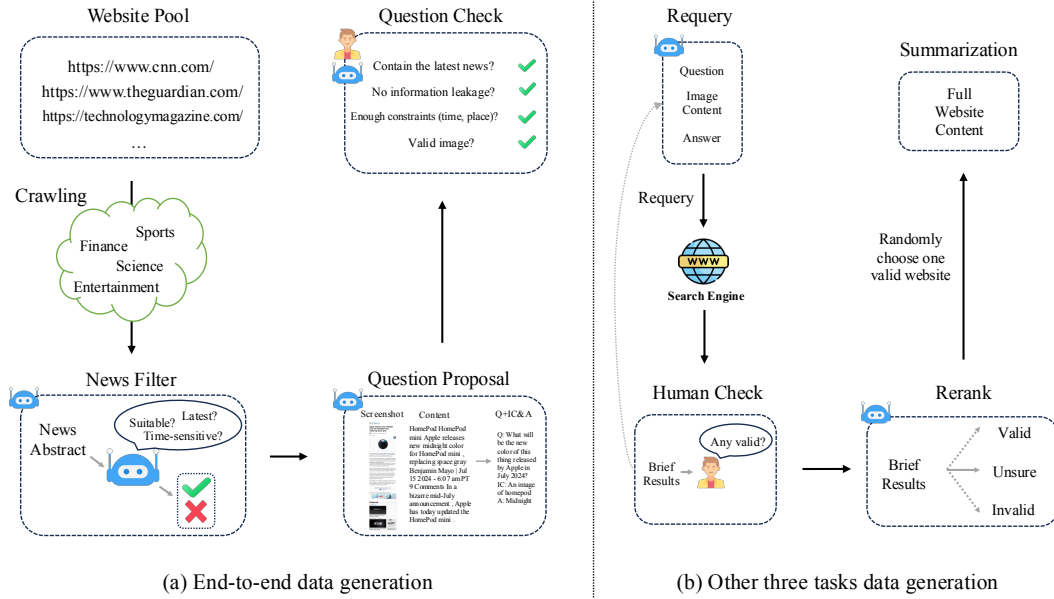


Figure 1: **Illustration of the Data Curation Pipeline.** We first collect the end-to-end data by crawling inside the website pool and prompting LMMs to raise questions based on the content. Then we further prompt the LMM to generate annotation for other three tasks. The human check is optional for the end-to-end data generation but compulsory for other three tasks data generation.

Now we introduce our automated/semi-automated data curation pipeline. The figure is shown in Fig. 1. We first define a website pool and a model pool. The website pool contains general news websites like CNN and expertise websites like Arxiv. The model pool contains state-of-art models for the data curation pipeline to guarantee diversity and fairness.

1. **End-to-end data curation.** We employ a web crawler to obtain all the sub-websites published later than a specific date. However, not all the websites are suitable for raising a question to test the LMMs’ searching capability. For example, some websites do not contain any recent news, while some websites’ contents are difficult to convert into a question with a definite answer. Therefore, we randomly choose a model from the model pool to serve as a news filter, prompting it to filter valid websites by providing some few-shot examples. Next, we provide the text contents and screenshots of the valid websites to a model from the model pool. The model is asked to raise several questions based on the website content. It is encouraged to raise questions unable to be answered only by text. As for the question with an image, the model is asked to briefly describe the image content, and we will later use the description to search in Bing and obtain the first result image. Finally, we apply the quality check of the generated questions and their corresponding images either by human or a model from the model pool.
2. **Rerank data generation.** We provide the generated requery to the search engine and retrieve K websites for rerank. Again, we provide the question, the question image content, and the answer to a model and ask to categorize each website into valid, unsure or invalid.
3. **Summarization data generation.** We randomly choose one website from the websites marked as valid in the last step and obtain its full content.

Table 1: **Evaluation Results of Different Complexity in MMSEARCH.** We categorize all the questions into three levels of complexity.

Model	All			End-to-end			Requery			Rerank			Summarization		
	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
<i>Closed-source LMMs with MMSEARCH-ENGINE</i>															
Claude 3.5 Sonnet (Anthropic, 2024)	56.2	53.6	48.9	53.0	49.8	45.0	49.3	37.6	35.0	80.3	84.5	75.3	59.8	59.9	58.3
GPT-4V (OpenAI, 2023)	55.3	56.4	53.1	52.7	52.9	50.2	52.7	41.4	38.8	76.5	82.8	80.2	55.0	63.9	54.3
GPT-4o (OpenAI, 2024b)	63.1	63.9	59.2	61.0	62.3	57.4	54.3	41.2	40.6	82.2	85.6	81.5	64.0	65.5	59.0
<i>Open-source LMMs with MMSEARCH-ENGINE</i>															
Mantis (Jiang et al., 2024a)	22.8	18.0	13.0	19.4	15.9	10.0	28.5	12.6	14.5	40.5	43.1	34.6	28.2	11.5	12.9
InternLM-XC2.5 (Zhang et al., 2024a)	26.6	19.3	18.1	28.5	20.0	16.7	24.3	26.0	25.0	0.0	0.0	0.0	39.5	30.3	42.7
InternLM-XC2.5 <sub>AnyRes</sub>	25.6	18.1	21.2	27.5	18.3	21.4	20.2	24.6	20.9	0.0	0.0	0.0	39.5	31.5	41.4
LLaVA-NeXT-Interleave (Li et al., 2024b)	33.2	25.6	23.3	28.5	20.2	16.9	31.0	22.1	22.7	61.4	51.7	49.4	41.5	41.6	45.0
mPlug-Owl3 (Ye et al., 2024)	37.1	28.5	27.7	30.5	20.6	19.5	40.8	28.3	23.9	77.7	73.0	70.4	44.8	44.1	48.4
mPlug-Owl3 <sub>AnyRes</sub>	37.5	32.8	29.0	31.4	26.4	21.5	38.6	26.9	26.0	76.5	77.6	67.9	44.0	39.6	48.3
InternVL2 (Chen et al., 2024c)	38.7	30.9	30.5	35.0	28.3	26.8	39.5	25.3	27.9	57.6	38.5	37.0	47.4	46.2	52.8
InternVL2 <sub>AnyRes</sub>	38.0	30.0	32.0	34.6	24.5	28.2	38.8	24.5	26.5	53.8	59.8	45.1	46.7	43.9	50.2
Idefics3 (Laurençon et al., 2024)	43.1	30.5	31.0	38.0	21.3	23.5	40.9	23.3	23.3	76.5	83.3	69.1	48.9	50.0	53.1
Idefics3 <sub>AnyRes</sub>	43.2	31.2	28.4	38.4	24.1	23.0	37.6	20.3	17.6	76.5	74.7	64.2	48.4	46.6	38.4
LLaVA-OneVision (Li et al., 2024a)	42.9	32.3	31.1	36.8	24.6	23.3	46.5	28.9	25.8	78.0	67.8	69.8	51.5	56.7	53.6
Qwen2-VL <sub>AnyRes</sub> (Qwen Team, 2024)	45.6	44.9	45.3	41.2	38.9	40.4	46.5	34.0	32.2	73.5	83.3	74.7	50.3	56.8	59.6
LLaVA-OneVision (72B)	52.2	48.0	49.0	47.6	41.6	44.0	51.0	39.6	33.2	79.5	85.1	83.3	60.3	63.7	60.6
Qwen2-VL <sub>AnyRes</sub> (72B)	53.7	53.4	50.2	50.7	48.6	46.9	52.3	40.0	37.1	72.3	82.2	77.8	58.5	67.0	53.6

Notably, human check is a must for the requery data generation process. There should be at least one valid website to guarantee the effectiveness of the generated requery. Only after human check of this step, the quality of rerank and summarization data generation is assured.

## C ADDITIONAL DATASET DETAILS

We manually annotate the complexity of the data based on the difficulty of the three steps in MMSearch-Engine:

1. **Requery difficulty.** This concerns the complexity of transforming the original question into an effective search query. Complex cases arise when the question references image content, requiring the LMM to first analyze the visual information and then synthesize it with the text question into a coherent search query. For instance, when a user asks about a landmark shown in an image, the LMM must first identify the landmark through image search and then incorporate this information into a text query about the landmark’s specific attributes.
2. **Rerank difficulty.** This dimension evaluates the challenge of identifying and prioritizing relevant search results. The difficulty primarily scales with the information scarcity. If there are only very limited websites containing useful information, it is more difficult to successfully retrieve and choose the website.
3. **Summarization difficulty.** This aspect involves both information synthesis and multi-modal reasoning challenges. In cases requiring synthesis, answers cannot be derived from a single source sentence - the LMM must integrate information scattered across different parts of the website. For example, comparing event frequencies across locations (like concert counts between cities) requires gathering and analyzing distributed data. Additionally, some questions demand analysis of both textual and visual website content, sometimes necessitating comparison with input images with images in the website.

Based on these criteria, we have categorized all questions into three difficulty levels, with the following distribution: hard (28%), medium (27.7%), and easy (44.3%). We provide the evaluation results grouped by the difficulty levels in Table 1.

## D FUTURE DIRECTION

Our proposed MMSEARCH-ENGINE can be enhanced through interactive user feedback loops. When the model produces an incorrect answer, users can identify the specific step where the error occurred and prompt the model to reconsider its reasoning. This iterative process allows for guided model refinement until accurate results are achieved. We conducted preliminary experiments on GPT-4o with this approach on three test cases. The results demonstrated that the LMM successfully interpreted user feedback and appropriately adjusted its responses based on the provided guidance. The model’s ability to understand the user input and revise its reasoning suggests significant potential for improving task accuracy.

## E ADDITIONAL EXPERIMENTS AND ANALYSIS

### E.1 SCALING TEST-TIME COMPUTE VS SCALING MODEL SIZE

Recent works such as OpenAI o1 (OpenAI, 2024a) and Li et al. (2024c) have highlighted the critical role of scaling test-time computation in enhancing model performance. Our end-to-end task, which requires multiple Internet interactions, presents an opportunity to investigate the potential of scaling test-time computation compared to scaling model size. To explore this, we conduct experiments using LLaVA-OneVision-7B (Li et al., 2024a), focusing on scaling test-time computation, and compare against LLaVA-OneVision-72B scaling in model size, which aims to provide insights into the relative benefits of increased inference computation versus increased model parameters.

For scaling up the test-time computation, we adopt a multi-modal search strategy similar to best-of-N solution, where ‘N’ denotes 25 in our settings. Specifically, for LLaVA-OneVision-7B, we first prompt the model to generate a requery 5 times, from which we selected the one with the highest requery score  $S_{req}$ . This requery is then used to retrieve brief results from 8 websites from a search engine. The model is again prompted 5 times to select the most informative website. After removing duplicates from the selected websites, we extract the full website content from the remaining ones and prompt the model to answer 5 times, obtaining 25 end-to-end outputs in total. We compute the F1 score for each answer against the ground truth and take the maximum as the model’s end-to-end score for the query. Table 2 shows that LLaVA-OneVision-7B (TTC) achieves the score of 55.2% in the end-to-end task, significantly enhancing the original score of 29.6%, which surpasses LLaVA-OneVision-72B’s 44.9% and GPT-4V’s 52.1%. This result reveals the substantial potential of scaling test-time computation, validating the effectiveness of this technique as introduced by OpenAI o1. Our findings provide valuable insights for future research in this domain, suggesting that increased inference computation may offer comparable or superior performance improvements to increased model size not only in math and code tasks, but also in multimodal search tasks.

Table 2: **Scaling Test-Time Compute vs Scaling Model Size.** ‘TTC’ and  $S_{e2e}$  denote Test-Time Computation and the score of end-to-end task.

Model	Inference Cost	$S_{e2e}$
LLaVA-OV-7B	1	29.6
LLaVA-OV-7B (TTC)	~25	55.2
LLaVA-OV-72B	~6	44.9

### E.2 DEFINITION OF ERRORS IN THE REQUERY AND SUMMARIZATION TASKS

Five types of requery error:

- *Lacking Specificity*, where the model fails to include all the specific information in the requery and therefore leads to sub-optimal search results. For example, the query is asking the release date of Vision Pro in China. However, the model omits the condition of China and directly asks about the release date of Vision Pro.
- *Inefficient Query*, where the model does not consider the real scenario and the requery is inefficient for the search engine to find the answer. For example, the query is asking whether the Van Gogh’s Sunflowers and Antoni Clavé’s Grand Collage are both oil paintings. Clearly, it is a commonsense that Van Gogh’s Sunflowers is an oil painting and Antoni Clavé’s Grand Collage is much less well-known. An efficient query should be asking about the images of Antoni Clavé’s Grand Collage and further determine if it is also an oil painting by directly looking at it. However, the model directly asks the original query to the



---

search engine. There is very little chance that an exact same question has ever been raised so probably this requery will bring very little helpful information.

- *Excluding Image Search Results*, where the model totally ignores the information in the screenshot of the image search results and therefore lacks important specific information in the requery. For example, the query is ‘When did this football player obtain the gold medal?’ and provides an image of the player. The model is supposed to find out the player’s name by viewing the image search result and raise a requery like ‘[PLAYER NAME] obtained the gold medal time’. However, the model fails to incorporate the player’s name in the requery and definitely the retrieved websites will not include any helpful information.
- *No Change*, where the model just uses the question as the query input to the search engine.
- *Irrelevant*, where the model either matches wrong information from the image search result or mistakenly understands the query and outputs an irrelevant requery.

Five types of summarization error:

- *Text Reasoning Error*, where the model fails to extract the answer from the website textual information.
- *Image-text Aggregation Error*, where obtaining the answer needs combining the information from both images and texts. The model fails to do so.
- *Image reasoning Error*, where the model fails to extract the answer from the image, and the answer can only be obtained from the image.
- *Hallucination* (Huang et al., 2023), where the model provides an unfaithful answer that cannot be grounded in the given content.
- *Informal*, the output format does not follow the prompt specifications, the same error type in the end-to-end task.

## F ADDITIONAL EXPERIMENTAL DETAILS

**Rationality of the weight in final score.** Considering the sequential nature of the tasks, the reason for our weighting scheme includes two complementary perspectives:

1. The importance of each task due to the cascaded nature is already reflected in the end-to-end score. Detailedly, although it is easy to discern that the upstreamed task is more important, it is difficult to assign a precise weight to each of them. So we do not manually assign the weights but directly focus mainly on the end-to-end score, which implicitly considers their cascaded nature.
2. The individual task weights serve as complementary metrics rather than indicators of relative importance. Relying solely on end-to-end evaluation, while comprehensive, may obscure the performance characteristics of individual components and hinder targeted improvements. We therefore maintain independent evaluation of each task, with the weight distribution designed to balance the prominence of the end-to-end metric against the component-level assessments. This dual evaluation strategy enables both system-level optimization and component-specific refinements.

**More Implementation Details** All our experiments of open-source models are conducted without any fine-tuning on search data or tasks. As for the prompts, the requery prompt contains 3 examples to better guide LMMs to output a valid requery. While prompts for other tasks are all in a zero-shot setting. We prompt the LMM to output as few words as possible for a better match with the ground truth. We employ the metric introduced in Section ?? . Besides, we recruit eight qualified college students and ask them to solve the problems in MMSEARCH independently, following the same pipeline of MMSEARCH-ENGINE. This score serves as a baseline for human performance. We conduct all experiments on NVIDIA A100 GPUs.

The input image dimensions for the webpage’s top section screenshot were set to  $1024 \times 1024$  pixels. For the full-page screenshot, we set the initial webpage width to 512 pixels, although the actual width of a small portion of webpages may vary due to its layout settings. Furthermore, considering that a

full-page screenshot can be extremely lengthy, directly inputting it as a single image into an LLM would result in excessive downsizing, making the content too vague for accurate identification. To address this, we segmented the full-page screenshot into multiple images, starting from the top, with each segment measuring 512 pixels in height. Because of the context length limitations of LMMs, the maximum number of full-page screenshot segments is therefore restricted to 10.

**Full-page Screenshot Slimming.** For the full-page screenshot, we compute the Sobel gradients (Kanopoulos et al., 1988) to detect the edges and generate a gradient magnitude image. We iteratively remove the areas with gradients below a threshold, which represent the blank areas. This approach, shown in Fig. 2, effectively reduces image size while maintaining the document content.



Figure 2: **Illustration of the Screenshot Slim Process.** We leverage Sobel gradients (Kanopoulos et al., 1988) to identify blank areas and remove them. After slimming, the screenshot size is largely reduced without any information loss.

**Model Sources.** For different LMMs, we select their latest models with size around 7B for evaluation to fully reveal their multimodal search proficiency. Table 3 presents the release time and model sources of LMMs used in MMSEARCH.

Table 3: **The Release Time and Model Source of LMMs Used in MMSEARCH.**

Model	Release Time	Source
GPT-4V (OpenAI, 2023)	2023-09	<a href="https://platform.openai.com/">https://platform.openai.com/</a>
GPT-4o (OpenAI, 2024b)	2024-05	<a href="https://platform.openai.com/">https://platform.openai.com/</a>
Claude 3.5 Sonnet (Anthropic, 2024)	2024-06	<a href="https://www.anthropic.com/news/claude-3-5-sonnet">https://www.anthropic.com/news/claude-3-5-sonnet</a>
InternLM-XC2.5 (Zhang et al., 2024a)	2024-07	<a href="https://github.com/InternLM/InternLM-XComposer">https://github.com/InternLM/InternLM-XComposer</a>
Mantis (Jiang et al., 2024a)	2024-05	<a href="https://tiger-ai-lab.github.io/Mantis/">https://tiger-ai-lab.github.io/Mantis/</a>
LLaVA-NeXT-Interleave (Li et al., 2024b)	2024-06	<a href="https://github.com/LLaVA-VL/LLaVA-NeXT">https://github.com/LLaVA-VL/LLaVA-NeXT</a>
InternVL2 (Chen et al., 2024c)	2024-07	<a href="https://github.com/OpenGVLab/InternVL">https://github.com/OpenGVLab/InternVL</a>
mPlug-Owl3 (Ye et al., 2024)	2024-08	<a href="https://github.com/X-PLUG/mPLUG-Owl">https://github.com/X-PLUG/mPLUG-Owl</a>
Idefics3 (Laurençon et al., 2024)	2024-08	<a href="https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3">https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3</a>
LLaVA-OneVision (Li et al., 2024a)	2024-08	<a href="https://llava-vl.github.io/blog/2024-08-05-llava-onevision/">https://llava-vl.github.io/blog/2024-08-05-llava-onevision/</a>
Qwen2-VL (Qwen Team, 2024)	2024-08	<a href="https://github.com/QwenLM/Qwen2-VL">https://github.com/QwenLM/Qwen2-VL</a>

**Input Prompts of LMM for Response Generation.** We showcase the input prompts of LMM for the three tasks respectively in Table 4-6. We adopt two types of prompts for queries with an image and without images. For query with an image, we specifically require the LMM to leverage the image search result to solve the task.

Table 4: **Input Prompt of LMMs for Requery.** We adopt two different prompts for the query with image input and without image input. We leverage a 3-shot prompt to guide the LMM to generate a reasonable requery.

Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question, which cannot be solved without external knowledge. Assume you have access to a text-only search engine (e.g., google). Please raise a query to the search engine to search for what is useful for you to answer the question correctly. Your query needs to consider the attribute of the query to search engine. Here are 3 examples:</p> <p>Question: Did Zheng Xiuwen wear a knee pad in the women’s singles tennis final in 2024 Paris Olympics? Query to the search engine: Images of Zheng Xiuwen in the women’s singles tennis final in 2024 Paris Olympics</p> <p>Question: When will Apple release iPhone16? Query to the search engine: iPhone 16 release date</p> <p>Question: Who will sing a French song at the Olympic Games closing ceremony? Query to the search engine: Singers at the Olympic Games closing ceremony, French song.</p> <p>Question: <i>{question}</i>.</p> <p>Query to the search engine (do not involve any explanation):</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image, which cannot be solved without external knowledge. Assume you have access to a search engine (e.g., google). Please raise a query to the search engine to search for what is useful for you to answer the question correctly. You need to consider the characteristics of asking questions to search engines when formulating your questions. You are also provided with the search result of the image in the question. You should leverage the image search result to raise the text query. Here are 3 examples:</p> <p>Question: Did Zheng Xiuwen wear a knee pad in the women’s singles tennis final in 2024 Paris Olympics? Query to the search engine: Images of Zheng Xiuwen in the women’s singles tennis final in 2024 Paris Olympics</p> <p>Question: When will Apple release iPhone16? Query to the search engine: iPhone 16 release date</p> <p>Question: Who will sing a French song at the Olympic Games closing ceremony? Query to the search engine: Singers at the Olympic Games closing ceremony, French song</p> <p>Question: <i>{query_image}{question}</i>. The image search result is: <i>{image_search_result}</i></p> <p>Query to the search engine (do not involve any explanation):</p>

Table 5: **Input Prompt of LMMs for Rerank.** We adopt two different prompts for the query with image input and without image input.

Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question and 8 website information related to the question (including the screenshot, snippet and title). You should now read the screenshots, snippets and titles. Select 1 website that is the most helpful for you to answer the question. Once you select it, the detailed content of them will be provided to help you correctly answer the question. The question is <math>\{question\}</math>. The website informations is <math>\{website\_information\}</math>. You should directly output 1 website’s index that can help you most, and enclose the website in angle brackets. The output format should be: &lt;Website Index &gt;. An example of the output is: &lt;Website 1 &gt;. Your answer:</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image. You are provided with the search result of the image in the question. And you are provided with 8 website information related to the question (including the screenshot, snippet, and title). You should now read the screenshots, snippets and titles of these websites. Select 1 website that is the most helpful for you to answer the question. Once you select it, the detailed content of them will be provided to help you correctly answer the question. The question is <math>\{query\_image\}\{question\}</math>. The image search result is <math>\{image\_search\_result\}</math>. The website information is <math>\{website\_information\}</math>. You should directly output 1 website’s index that can help you most, and enclose the website in angle brackets. The output format should be: &lt;Website Index &gt;. An example of the output is: &lt;Website 1 &gt;. Your answer:</p>

Table 6: **Input Prompt of LMMs for Summarization.** We adopt two different prompts for the query with image input and without image input.

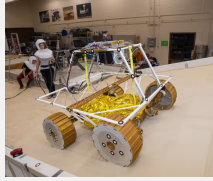
Question	Prompt
Query without image	<p>You are a helpful assistant. I am giving you a question and 1 website information related to the question. Please follow these guidelines when formulating your answer: 1. If the question contains a false premise or assumption, answer “invalid question”. 2. When answering questions about dates, use the yyyy-mm-dd format. 3. Answer the question with as few words as you can.</p> <p>You should now read the information of the website and answer the question. The website information is <i>{website_information}</i>. The question is <i>{question}</i>. Please directly output the answer without any explanation:</p>
Query with image	<p>You are a helpful assistant. I am giving you a question including an image. You are provided with the search result of the image in the question. And you are provided with 1 website information related to the question. Please follow these guidelines when formulating your answer: 1. If the question contains a false premise or assumption, answer “invalid question”. 2. When answering questions about dates, use the yyyy-mm-dd format. 3. Answer the question with as few words as you can.</p> <p>You should now read the information of the website and answer the question. The website information is <i>{website_information}</i>. The image search result is <i>{image_search_result}</i>. The question is <i>{query_image}{question}</i>. Please directly output the answer without any explanation:</p>

## G MORE DATA DETAILS

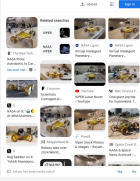
### G.1 DATA EXAMPLE OF 4 EVALUATION TASKS

**Query Information**

**Image**



**Image Search Result**



**Question:**  
In which day of 2024 was the lunar rover project in the picture announced to be canceled?

**Task1 Requery**

Query Information

**LMM Requery:**  
VIPER was announced to be cancelled

**Requery Annotation:**  
NASA Ends VIPER Project

**Task2 Rerank**

Query Information

Brief Result

**<Website 1>:**  
**Title:** NASA Ends VIPER Project, Continues Moon Exploration  
**Snippet:** NASA Ends VIPER Project, Continues Moon Exploration. Tiernan P. Doyle. Jul 17, 2024. RELEASE 24-094. ... (Volatiles Investigating Polar Exploration Rover) project. NASA stated cost increases, delays to the launch date, and the risks of future cost growth as the reasons to stand down on the mission. The rover was originally planned to launch in ...

**<Website 2>:**  
**Title:** NASA cancels VIPER lunar rover - SpaceNews  
**Snippet:** The VIPER lunar rover. Credit: NASA. BUSAN, South Korea 'u2014 NASA has canceled a robotic lunar rover mission that would have searched for ice at the south pole of the moon, citing development ...

**<Website 3>:**  
**Title:** NASA Ends VIPER Project, Continues Moon Exploration  
- Yahoo Finance  
**Snippet:** NASA Ends VIPER Project, Continues Moon Exploration. PR Newswire. Wed, Jul 17, 2024, 4:13 PM 3 min read. Link Copied. 0. WASHINGTON, July 17, 2024 /PRNewswire/ -- Following a comprehensive ...

**<Website 4>:**  
**Title:** NASA axes robotic lunar rover project VIPER due to rising costs - MSN  
**Snippet:** NASA has ended its VIPER project, which was hoping to launch the agency's first robotic lunar rover to the moon, due to the increasing costs of the program. "Decisions, of course, like this are ...

**LMM Rerank:**  
<Website 2>


**Rerank Annotation:**  
**Valid:** [< Website 1>, <Website 2>, <Website 3>]  
**Unsure:** [<Website 4>, <Website 5>]  
**Invalid:** [<Website 6>, <Website 7>, <Website 8>]

**Task3 Summarization**

Query Information

Full Website Content

**Full-page Screenshot**



**Content:**  
would have searched for ice at the south pole of the moon , citing development delays and cost overruns . NASA announced July 17 that it would end development of the Volatiles Investigating Polar Exploration Rover ( VIPER ) mission . The rover , to be sent to the south polar region of moon on a commercial lander called Griffin from Astrobotic Technology , would have explored terrain including permanently shadowed regions to better understand the extent and form of water ice there . At a briefing to announce the cancellation , agency officials said costs of VIPER had grown  
Posted inCivil NASA cancels VIPER lunar rover by Jeff Foust July 17 , 2024July 17 , 2024 Click to share on X ( Opens in new window ) Click to share on Facebook ( Opens in new window ) Click to share on LinkedIn ( Opens in new window ) Click to share on Reddit ( Opens in new window ) Click to email a link to a friend ( Opens in new window ) Click to share on Clipboard ( Opens in new window ) BUSAN , South Korea — NASA has canceled a robotic lunar rover mission that 's lander that would deliver the rover to the moon under a CLPS task order worth \$ 322 million . NASA said Griffin was now expected to be ready for the mission no earlier than September 2025 . With VIPER canceled , NASA will retain the task order for Griffin . The mission will instead become a technology demonstrator , carrying a mass simulator in place of the rover to test Griffin 's ability to land large payloads . Kearns said NASA considered flying science payloads instead , but since the lander was designed for carrying a rover by more than 30 % , triggering a termination review by the agency . NASA had confirmed VIPER in 2021 at a cost of \$ 433.5 million . Joel Kearns , deputy associate administrator for exploration in NASA 's Science Mission Directorate , said the latest estimate was \$ 609.6 ...

**LMM Answer:**  
July 17

**Answer Annotation:**  
07-17

**Task4 End-to-end**

Query Information

**LMM Answer:**  
Sep 12

**Answer Annotation:**  
07-17

Figure 3: **Example Input, LMM Output, and Ground Truth for Four Evaluation Tasks.** The color-coding of each module corresponds to Fig. ?? . Task1 Requery (green), Task2 Rerank (purple), Task3 Summarization (blue), and Task4 End-to-end (yellow) are shown. Image best viewed in color.



---

## G.2 SUBFIELD DEFINITION

**News** area encompasses a vast spectrum of information, ranging from everyday events to engaging entertainment content and specialized fields such as scientific discoveries and financial analysis. This comprehensive coverage serves as a rigorous assessment of the model’s ability to process information in diverse domains. We divide this expansive area into eight distinct subfields:

- **Traditional Sports:** Data concerning traditional athletic competitions, team performances, player statistics, and sporting events. This includes scores, league standings, player transfers, and analysis of various professional sports across different leagues and countries.
- **e-Sports:** Information about competitive video gaming, including tournament results, player rankings, and league information. This covers various game titles, team formations, streaming viewership statistics, and tournament information.
- **Technology:** Information about technological innovations, gadgets, software developments, and tech industry news. This includes product launches, software updates, cybersecurity issues, and artificial intelligence advancements.
- **Paper:** Content related to academic papers, research publications, and scholarly articles in various artificial intelligence fields. The queries include method explanation, figure understanding, and experiment settings.
- **Entertainment:** Data about movies, music, television, celebrities, and other forms of popular entertainment. It also includes data concerning video games.
- **Finance:** Information on financial markets, economic indicators, business news, and monetary policies. This covers stock prices, company earnings reports, company financial statements, and regulatory news regarding finance.
- **General News:** Broad coverage of various news topics not specific to any particular subfield. This includes a mix of local and global events, human interest stories, lifestyle articles, climate news, and general interest content that doesn’t fit neatly into other specialized news subfields.
- **False Premise:** Data related to misinformation or incorrect assumptions in the query. This subfield focuses on fact-checking capabilities. All the answers to the queries of this subfield are ‘invalid question’.

**Knowledge** area represents broad subfields of information and data related to general knowledge across various disciplines. This area concentrates on rare knowledge that most LMMs fail to answer. We categorize this area into five subfields:

- **Architecture:** Information about building design, architectural styles, building information, and construction projects. This includes city landmarks, the comparison of architectural styles, and multi-view architecture matchings.
- **Arts:** Data concerning visual arts, drawings, sculptures, badges, and other forms of creative expression. This covers artwork details, artist profiles, artwork history, and artwork style comparisons.
- **Fashion:** Content related to clothing trends, fashion brands, and designer collections. This includes retail price, clothing style, release date, and brand information.
- **Astronomy:** Information about celestial objects, space exploration, astronomical phenomena, and related research. This covers observational data from telescopes and image results from space missions. The questions focus on the background information of these celestial objects presented in the query image.
- **Anime:** Data about Japanese animation, including series storylines and character information. This encompasses character background, character appearance, voice actor information, and chapter information.
- **Auto:** Content related to automobiles, including vehicle specifications, industry trends, and automotive technology. This covers new car models, performance test results, coefficients of cars, and release date.

## H QUALITATIVE EXAMPLES

**Question:** In the LPL 2024 Summer season, how many teams were in Group Ascend?

**Answer:** 9

**Area:** News

**Subfield:** ⚡ Sports

Round1 Requery

**GPT-4o:** LPL 2024 Summer season Group Ascend teams count

**Qwen2-VL:** LPL 2024 Summer season Group Ascend teams

**LLaVA-OneVision:** LPL 2024 Summer season Group Ascend teams

Round2 Rerank

**GPT-4o Brief Results:**



**<Website 1>:**  
**Title:** LPL 2024 Summer - Leaguepedia | League of Legends Esports Wiki - Fandom  
**Snippet:** The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of Chinas professional League of Legends league. The summer season sees major changes of LPL structure, including introduction of "Fearless Draft" rules and return of group stages. Seventeen teams play against each other first in four double round robins, and then in two single round robins.



**<Website 5>:**  
**Title:** Group Stage / LPL 2024 Summer - schedule, results | 2014 Escorenews  
**Snippet:** LPL 2024 Summer Match results, calendar, VODs, stream, team rosters, schedule ... VODs, stream, team rosters, schedule. LoL News Bets and predictions Matches Events Teams Players. LPL 2024 Summer LoL | 2022 2024-06-01 - 0000-00-00 | 2022 \$578600. Stats. Event Placements Qualifier Group Stage Playoff Regional Finals. Group Stage Playoff - Phase 3 ...



**<Website 2>:**  
**Title:** LPL Summer 2024 - Liqupid League of Legends Wiki  
**Snippet:** The LPL Summer 2024 split is the second split of the 2024 LPL season. The league maintains all 17 teams from the Spring Split, and will be held across China, in cities such as Shanghai, Suzhou, Shenzhen, Xi'an and Beijing. Bilibili Gaming is the defending title champion. This split, the LPL will experiment with a different format, featuring a ...



**<Website 6>:**  
**Title:** 2024 LPL season - Wikipedia  
**Snippet:** The 2024 LPL season is the 12th and ongoing season of the League of Legends Pro League ... The bottom two ascend group teams and the top four nirvana group teams will have to contest an additional match. [6] As per usual, the champion for Summer 2024 will qualify for the 2024 World Championship as China's number one seed. Spring. Regular Season



**<Website 3>:**  
**Title:** LPL 2024 Summer Placements - Leaguepedia | League of Legends Esports Wiki  
**Snippet:** The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of Chinas professional League of Legends league. The summer season sees major changes of LPL structure, including introduction of "Fearless Draft" rules and return of group stages. Seventeen teams play against each other first in four double round robins, and then in two single round robins.



**<Website 7>:**  
**Title:** JD Gaming vs. EDward Gaming / LPL 2024 Summer Placements - Reddit  
**Snippet:** JD Gaming vs. EDward Gaming / LPL 2024 Summer Placements - Week 4 - Group A / Post-Match Discussion LPL 2024 SUMMER ... With this win by JDG, FPX advance to Group Ascend alongside JDG for the LPL 2024 Summer Season. JDG | Leaguepedia | Liqupidia | Website | Twitter EDG | Leaguepedia | Liqupidia ... Team WE vs. Bilibili Gaming / LPL 2024 ...



**<Website 4>:**  
**Title:** LPL 2024 Summer LoL Coverage | GosuGamers  
**Snippet:** The 2024 League of Legends Pro League (LPL) is the second split of the twelfth year of China's professional League of Legends league. The summer season sees major changes of LPL structure. Placements tournament will determine the groupings for the group stage. Jun 2024. 01.



**<Website 8>:**  
**Title:** League of Legends LPL 2024 Summer Split - Sportskeeda  
**Snippet:** The LPL 2024 Summer Split format is vastly different, and a big change for a major region. The first stage of the Summer Split will be the Placements stage. Teams are divided into four groups with ...

**GPT-4o Rerank:** <Website 6>

Figure 4: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

## REFERENCES

- Anthropic. Claude-3.5. <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

- 
- Patrice B  chard and Orlando Marquez Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*, 2024.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023a.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024a.
- Jun Chen, Deyao Zhu<sup>1</sup> Xiaoqian Shen<sup>1</sup> Xiang Li, Zechun Liu<sup>2</sup> Pengchuan Zhang, Raghuraman Krishnamoorthi<sup>2</sup> Vikas Chandra<sup>2</sup> Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023b.
- Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*, 2024b.
- 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*, 2024c.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *ICML 2024*, 2024.
- Zilu Guo, Hongbin Lin, Zhihao Yuan, Chaoda Zheng, Pengshuo Qiu, Dongzhi Jiang, Renrui Zhang, Chun-Mei Feng, and Zhen Li. Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. *arXiv preprint arXiv:2503.10529*, 2025a.

- 
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.
- Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025b.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.
- Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3170–3180, 2021.
- Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.
- Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, et al. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. *arXiv preprint arXiv:2411.18623*, 2024.
- Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024a.
- Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024b.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.
- Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.
- Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

- 
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 2024c.
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *ECCV 2024*, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.
- 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 2024a. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. *NeurIPS 2024*, 2024b.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.
- Feipeng Ma, Hongwei Xue, Yizhou Zhou, Guangting Wang, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Visual perception by large language model’s weights. *Advances in Neural Information Processing Systems*, 37:28615–28635, 2025.
- OpenAI. GPT-4V(ision) system card, 2023. URL <https://openai.com/research/gpt-4v-system-card>.
- OpenAI. Openai o1. [Online], 2024a. <https://openai.com/index/learning-to-reason-with-llms/>.
- OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024b.
- Qwen Team. Qwen2-vl. 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL <https://api.semanticscholar.org/CorpusID:231591445>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Yiwen Tang, Zoey Guo, Zhuhao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Xuelong Li, et al. Exploring the potential of encoder-free architectures in 3d lmm. *arXiv preprint arXiv:2502.09620*, 2025.

- 
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Peng Xia, Ze Chen, Juanxi Tian, Gong Yangrui, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. *arXiv preprint arXiv:2410.10139*, 2024a.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024b.
- Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1093, 2024c.
- Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. Crag-comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*, 2024.
- Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023a.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv, abs/2308.02490*, 2023.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.
- Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.



- 
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=d4UiXAHN2W>.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024c.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. *arXiv preprint arXiv:2412.09618*, 2024a.
- Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024b.

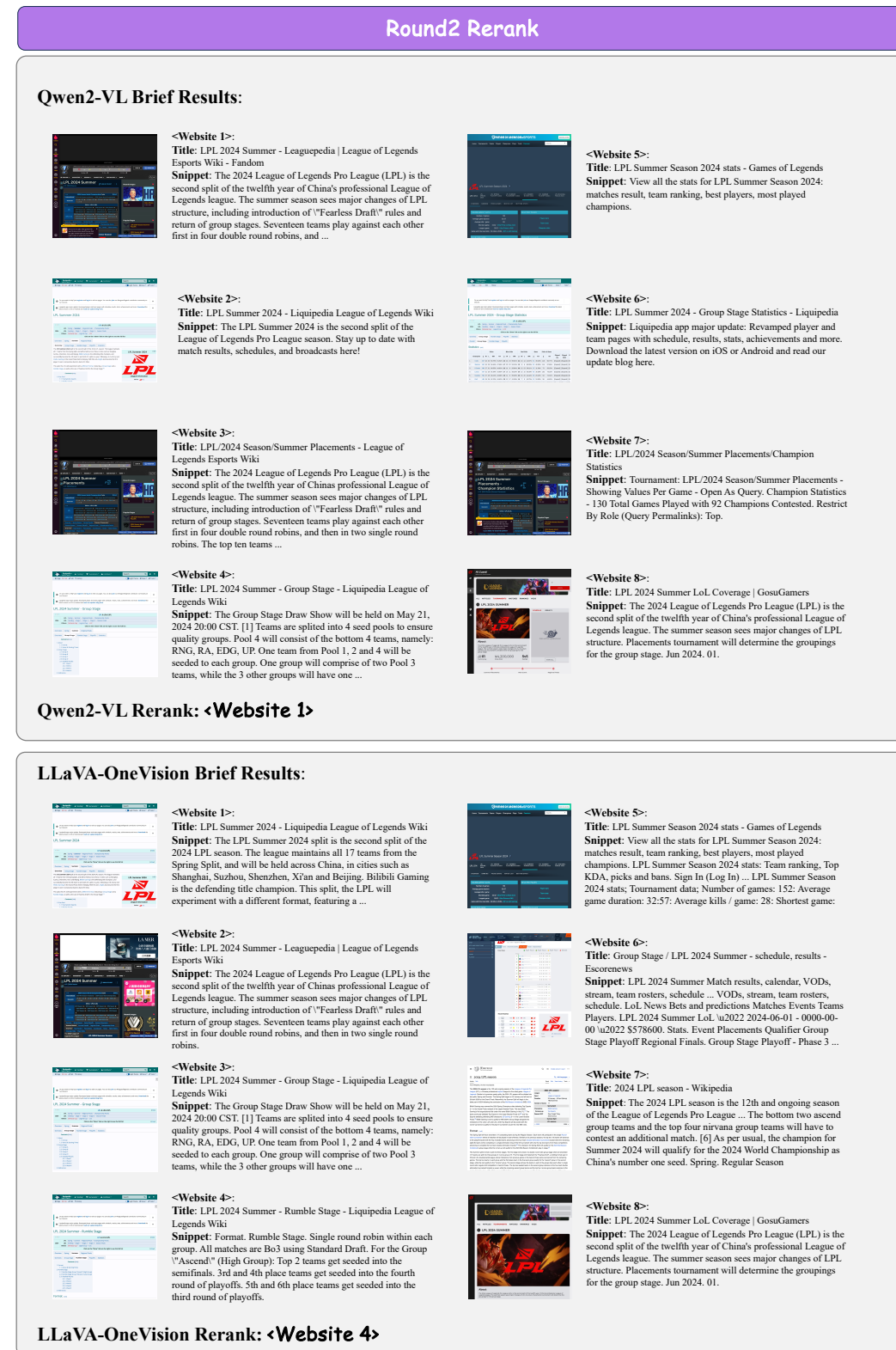
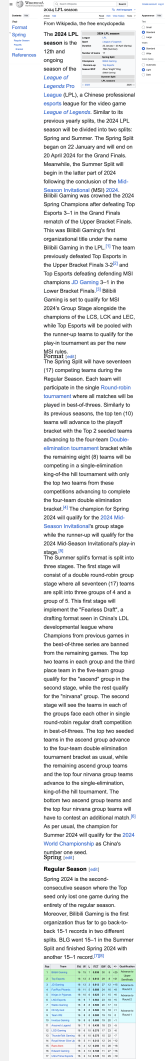


Figure 5: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

## Round3 Summarization

### GPT-4o

#### Full-page Screenshot:



#### Content:

for Spring 2024 will qualify for the 2024 Mid-Season Invitational 's group stage while the runner-up will qualify for the 2024 Mid-Season Invitational 's play-in stage . [ 5 ] The Summer split 's format is split into three stages . The first stage will consist of a double round-robin group stage where all seventeen ( 17 ) teams are split into three groups of 4 and a group of 5 . This first stage will implement the " Fearless Draft " , a drafting format seen in China 's LDL developmental league where Champions from previous games in the top four nirvana group teams advance to the single-elimination , king-of-the hill tournament . The bottom two ascend group teams and the top four nirvana group teams will have to contest an additional match . [ 6 ] As per usual , the champion for Summer 2024 will qualify for the 2024 World Championship as China 's number one seed . Spring [ edit ] Regular Season [ edit ] Spring 2024 is the second-consecutive season where the Top seed only lost one game during the entirety of the regular season . Moreover , Bilibili Gaming is the first organization 2024 LPL season Add links From Wikipedia , the free encyclopedia Sports season 2024 LPL season League LPL Sport League of Legends Duration 22 January – 20 April ( Spring ) TBD ( Summer ) Number of teams 17 Spring Split Champions Bilibili Gaming Runners-up Top Esports Season MVP Zhuo " knight " Ding ( Bilibili Gaming ) Summer Split LPL seasons ← 2023 2025 → The 2024 LPL season is the 12th and ongoing season of the League of Legends Pro League ( LPL ) , a Chinese professional esports league for the video game League of Legends .

Similar to its previous yearly splits , the 2024 LPL season will be divided into two splits : Spring and Summer . The Spring Split began on 22 January and will end on 20 April 2024 for the Grand Finals . Meanwhile , the Summer Split will begin in the latter part of 2024 following the conclusion of the Mid-Season Invitational ( MSI ) 2024 . Bilibili Gaming was crowned the 2024 Spring Champions after defeating Top Esports 3–1 in the Grand Finals rematch of the Upper Bracket Finals . This was Bilibili Gaming 's first organizational title under the name Bilibili Gaming in the LPL . [ 1 ] The team previously defeated Top Esports in the Upper Bracket Finals 3–2 [ 2 ] and Top Esports defeating defending MSI champions JD Gaming 3–1 in the Lower Bracket Finals . [ 3 ] Bilibili Gaming is set to qualify for MSI 2024 's Group Stage alongside the champions of the LCS , LCK and LEC , while Top Esports will be pooled with the runner-up teams to qualify for the play-in tournament as per the new MSI rules . Format [ edit ] The Spring Split will have seventeen

thus far to go back-to-back 15-1 records in two different splits . BLG went 15–1 in the Summer Split and finished Spring 2024 with another 15–1 record . [ 7 ] [ 8 ] Pos Team Pld W L PCT GW GL +/- Qualification 1 Bilibili Gaming 16 15 1 0.938 30 5 +25 Advance to Upper Semifinals 2 Top Esports 16 13 3 0.813 29 8 +21 3 JD Gaming 16 13 3 0.813 27 12 +15 Advance to Round 3 4 FunPlus Phoenix 16 11 5 0.688 24 16 +8 5 Ninjas in Pyjamas 16 10 6 0.625 Gaming will return to MSI 2024 as one of the four group-stage qualifying teams . Top Esports ' defeat means their qualification to MSI 2024 will be through the Play-In stage where they are group alongside the 2023 League of Legends World Champions T1 . [ 10 ] Bracket [ edit ] Round 1 Round 2 Round 3 Upper semifinals Upper final Final 1 Bilibili Gaming 3 4 FunPlus Phoenix 1 5 Ninjas in Pyjamas 1 5 Ninjas in Pyjamas 3 5 Ninjas in Pyjamas 3 1 Bilibili Gaming 3 8 Oh My God 1 9 Team WE 2 2

April 2024 . ^ " All teams qualified for MSI 2024 League of Legends | ONE Esports " . www.oneesports.gg . 20 April 2024 , Retrieved 20 April 2024 . v t e League of Legends Pro League Teams Anyone 's Legend Bilibili Gaming Edward Gaming FunPlus Phoenix Invictus Gaming JD Gaming LGD Gaming LNG Esports Ninjas in Pyjamas Oh My God Rare Atom Royal Never Give Up Team WE Top Esports TT Gaming Ultra Prime Weibo Gaming Seasons 2020 2021 2022 2023 2024 v t e 2024 in professional League of Legends competition International Mid-Season Invitational Esports World Cup World Championship Regional LCS LEC LCK LPL PCS Retrieved from " https : //en.wikipedia.org/w/inde x.php ? title=2024\_LPL\_season & oldid=1237480104 " Categories : League of Legends Pro League seasons 2024 in esports Hidden categories : Use dmy dates from April 2024 Articles with short description Short description matches Wikidata

GPT-4o Summarize: 2



Figure 6: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

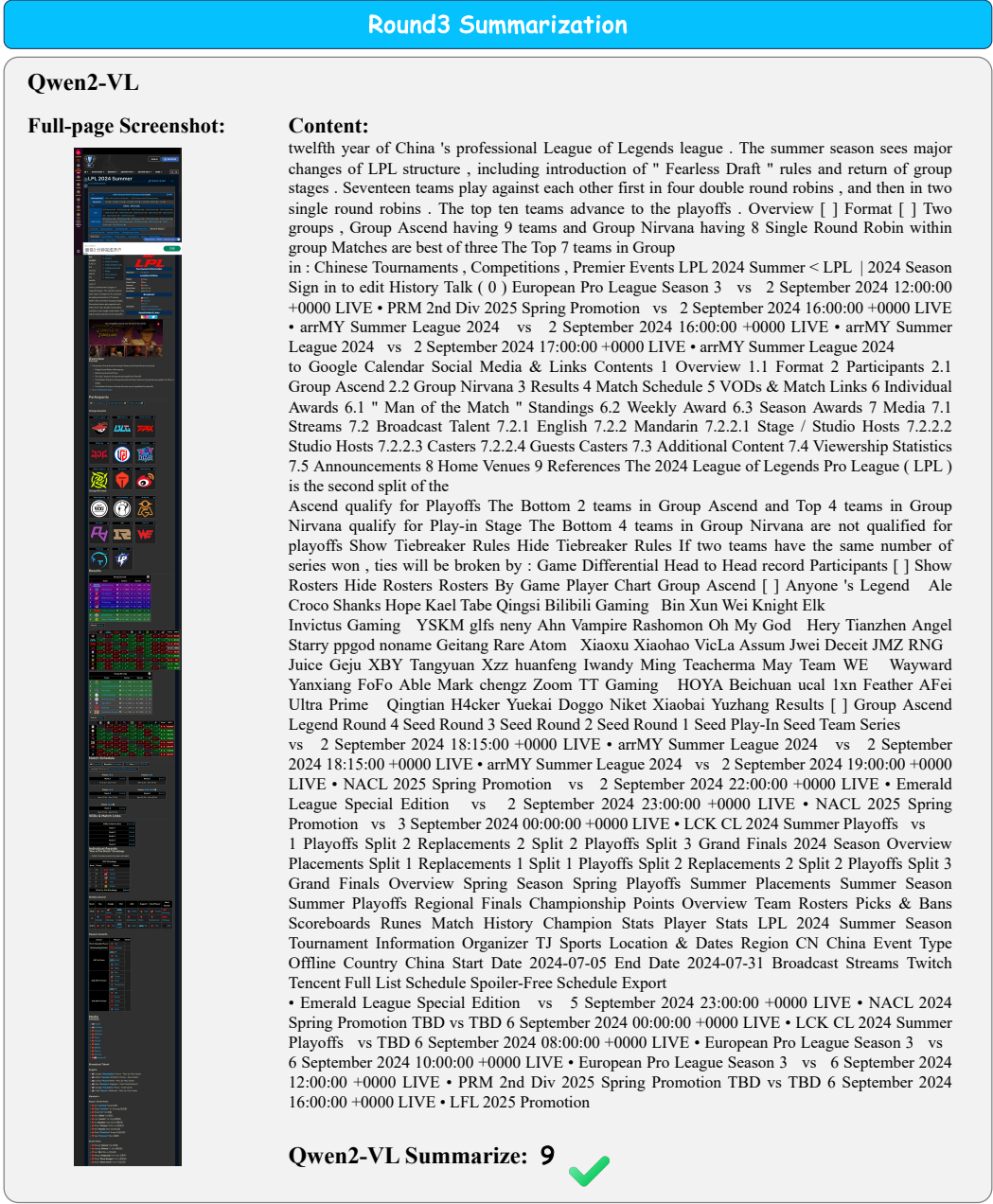
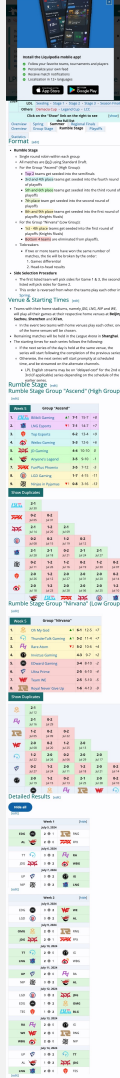


Figure 7: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

## Round3 Summarization

### LLaVA-OneVision

#### Full-page Screenshot:



#### Content:

register and log in to edit our pages . You can also join our # leagueoflegends contributor community on our Discord . Liquipedia app match pages updated ! Liquipedia app match pages are overhauled ! Download on Android or iOS ! Liquipedia app 's match pages got completely revamped with game data , standings , VODs and more ! Download the the latest version on iOS or Android and read our update blog here . LPL 2024 Summer - Rumble Stage From Liquipedia League of Legends Wiki < LPL | 2024/Summer v Group " Ascend " ( High Group ) 2.2 Rumble Stage Group " Nirvana " ( Low Group ) 2.3 Detailed Results 2.3.1 Week 1 2.3.2 Week 2 2.3.3 Week 3 2.3.4 Week 4 2.3.5 Week 5 3 References Format [ edit ] Rumble Stage Single round robin within each group All matches are Bo3 using Standard Draft . For the Group " Ascend " ( High Group ) : Top 2 teams get seeded into the semifinals 3rd and 4th place teams get seeded into the fourth round of playoffs 5th and 6th place teams get seeded into the d e LPL & LDL/LSPL 2024 LPL Group " Ascend " ( High Group ) [ edit ] Group " Ascend " Week 5 Week 1 Week 2 Week 3 Week 4 Week 5 1 . Anyone 's Legend 1-0 2-0 +2 1 . LNG Esports 1-0 2-0 +2 1 . Weibo Gaming 1-0 2-0 +2 Bilibili Gaming 0-0 0-0 0 LGD Gaming 0-0 0-0 0 Top Esports 0-0 0-0 0 7 . FunPlus Phoenix 0-1 0-2 -2 7 . JD Gaming 0-1 0-2 -2 7 . Ninjas in Pyjamas 0-1 0-2 -2 1 . LNG Esports 3-0 6-1 +5 2 . Anyone 's Legend ▼ 1 3-1 2024 - 17:00 CST Match Page 36:33 22:46 MVPs : Bin , knight Bans Game 1 Game 2 JDG 0 2 TES JD Gaming 0 2 Top Esports July 31 , 2024 - 19:00 CST Match Page 41:08 26:12 MVPs : 369 , 369 Bans Game 1 Game 2 Match was held in Beijing . References [ edit ] Retrieved from " https : //liquipedia.net/leagueoflegends/index.php ? title=LPL/2024/Summer/Rum ble Stage & oldid=828570 " Hidden categories : Pages reading from original match table Pages storing into original game table Pages storing into original match table Do you want to help ? Just LNG LGD Gaming 1 2 LNG Esports July 27 , 2024 - 15:10 CST Match Page 32:40 30:05 30:02 MVPs : haichao , Hang , Weiwei Bans Game 1 Game 2 Game 3 Match was held in Suzhou . EDG 2 0 UP EDward Gaming 2 0 Ultra Prime July 27 , 2024 - 17:10 CST Match Page 29:19 34:56 MVPs : Cryin , Jiejie Bans Game 1 Game 2 FPX 0 2 TES FunPlus Phoenix 0 2 Top Esports July 27 , 2024 - 19:30 CST Match Page 33:43 24:02 MVPs : 19:10 CST Match Page 26:56 27:10 MVPs : Zika , Weiwei Bans Game 1 Game 2 Match was held in Shenzhen . Week 2 [ edit ] Week 2 July 8 , 2024 EDG 0 2 WE EDward Gaming 0 2 Team WE July 8 , 2024 - 17:10 CST Match Page 27:19 29:46 MVPs : Yanxiang , Able Bans Game 1 Game 2 Match was held in Xi'an . LGD 0 2 AL LGD Gaming 0 2 Anyone 's Legend July 8 , 2024 - 19:10 CST Match Page 28:09 28:36 MVPs , 2024 - 17:10 CST Match Page Spring Summer Regional Finals Championship Points LDL Seeding Stage 1 Stage 2 Stage 3 Season Finals Others Demacia Cup Legend Cup LCC Click on the " Show " link on the right to see the full list 2023 LPL Spring Summer Regional Finals Championship Points LDL Stage 1 Stage 2 Stage 3 Others Demacia Cup LCC 2022 LPL Spring Summer Regional Finals Championship Points LDL Spring Summer Others Demacia Cup LCC 2021 LPL Spring Summer Regional Finals Championship Points LDL Spring Summer Others Demacia Cup LCC 2020 LPL Spring Summer Regional Finals Championship LNG Esports July 7 , 2024 - 33:22 41:37 MVPs : Xiaohao , Xiaoxu Bans Game 1 Game 2 JDG 0 2 WBG JD Gaming 0 2 Weibo Gaming July 6 , 2024 - 19:10 CST Match Page 43:33 33:25 MVPs : Tarzan , Breathe Bans Game 1 Game 2 Match was held in Beijing . July 7 , 2024 UP 0 2 iG Ultra Prime 0 2 Invictus Gaming July 7 , 2024 - 17:10 CST Match Page 31:35 39:02 MVPs : Glfs , Ahn Bans Game 1 Game 2 NIP 0 2 LNG Ninjas in Pyjamas 0 2 July 27 , 2024 LGD 1 2

LLaVA-OneVision Summarize: 8 ✖

Figure 8: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

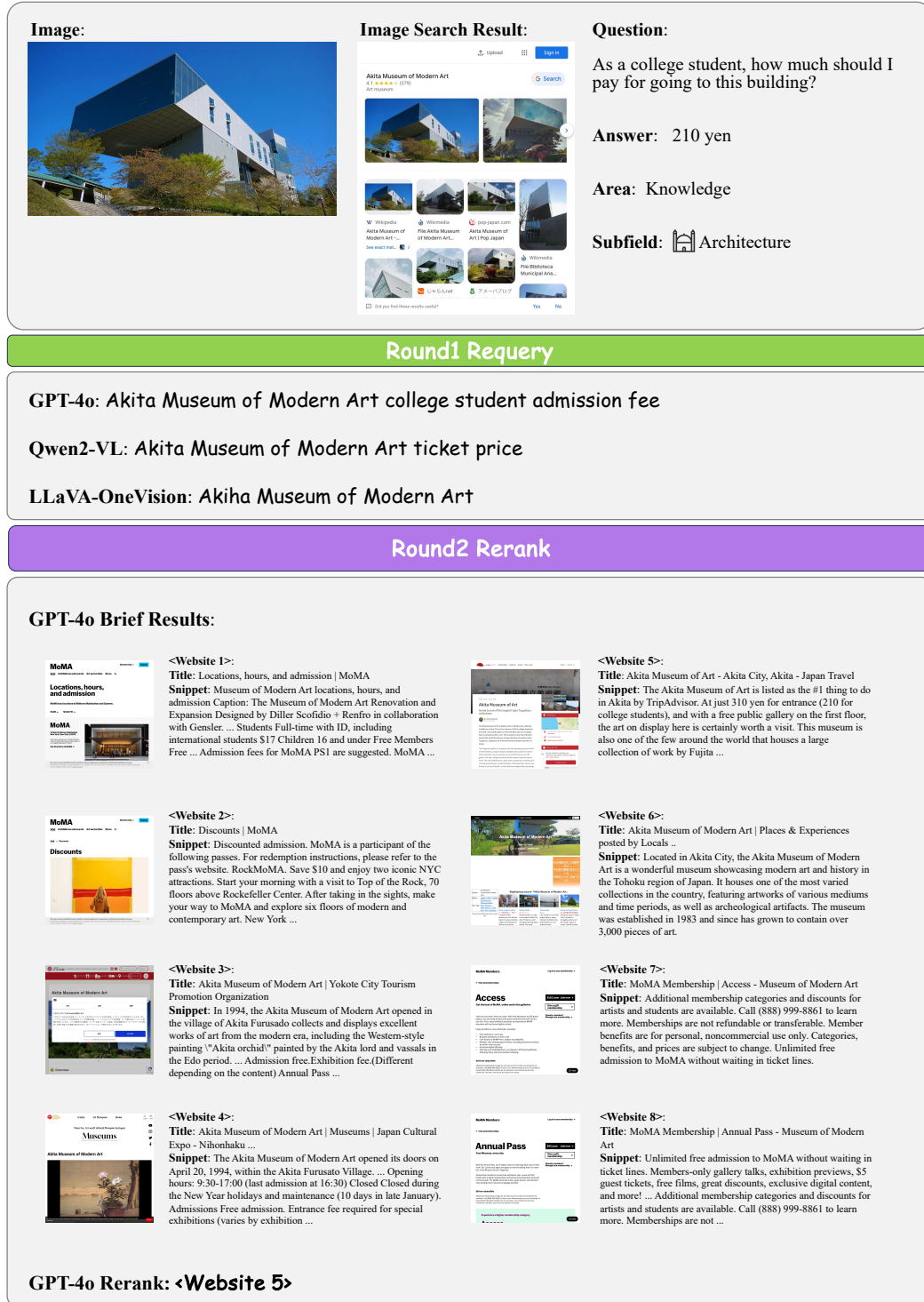
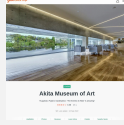


Figure 9: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.



## Round2 Rerank

### Qwen2-VL Brief Results:



**<Website 1>**  
**Title:** Akita Museum of Art - Must-See, Access, Hours & Price  
**Snippet:** This art museum was opened in 2013 in "Area Nakaiichi" near Akita Station as a hub of arts and cultures. The modern building was designed by Tadao Ando. The interior with repeated triangles and inorganic concrete walls creates out-of-the-ordinary artistic space.



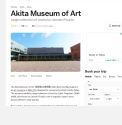
**<Website 2>**  
**Title:** Akita Museum of Modern Art | Yokote City Tourism Promotion Organization  
**Snippet:** In 1994, the Akita Museum of Modern Art opened in the village of Akita Furusado collects and displays excellent works of art from the modern era, including the Western-style painting "Akita orchid" painted by the Akita lord and vassals in the Edo period. I will.



**<Website 3>**  
**Title:** Akita Museum of Modern Art Opening hours, specific address, ticket ...  
**Snippet:** How does Akita Museum of Modern Art play? Including Akita Museum of Modern Art ticket information, prices and fees, reservations, opening hours, transportation, maps, addresses, travel guides, recommended accommodation, nearby restaurants and other practical information, book Akita Museum of Modern Art tickets at super-valuable prices through Trip com to enjoy Super discounts or learn more ...



**<Website 4>**  
**Title:** Akita Museum of Modern Art Details / Explore Japan Travel by NAVITIME ...  
**Snippet:** This museum was opened in 1994 in Yokote City, Akita Prefecture. It primarily exhibits modern works by Akita native or related artists. In addition, the Hi-Vision Gallery holds the museum's collection. The first floor has a Museum School offering regular lectures on art and many people participate.



**<Website 5>**  
**Title:** Akita Museum of Art - Akita Travel - japan-guide.com  
**Snippet:** The Akita Museum of Art (u79cbu7530u770c u7acb u78e u8853 u9928, Akita Kenritsu Bijutsukan) is an art museum in Akita City designed by renowned architect Ando Tadao. The museum exhibits a large collection of work by Fujita Tsuguharu (1886-1968), also known as Leonard Foujita, who is arguably Japan's most famous Western style painter.



**<Website 6>**  
**Title:** Akita Museum of Modern Art - Wikipedia  
**Snippet:** Akita Museum of Modern Art (u79cbu7530u770c u7acb u8fd1 u4ec3 u78e u8853 u9928, Akita Kenritsu Kindai Bijutsukan) opened in Yokote, Akita Prefecture, Japan in 1994 and houses an important collection of Akita ranga .



**<Website 7>**  
**Title:** Akita Museum of Modern Art | Places & Experiences posted by Locals ...  
**Snippet:** The Akita Museum of Modern Art is open from 10am until 5pm every day, and admission is free. Whether you're an art enthusiast or a casual visitor, the museum is an excellent option for exploring the history and cultural significance of Akita's art scene. [Text: @AI, Translation: @AI]



**<Website 8>**  
**Title:** Akita Museum of Art - Travel in Japan  
**Snippet:** The Akita Museum of Art in Akita City, Japan, showcases modern and contemporary Japanese art, including over 5,000 pieces. Highlights include works by Tsuguharu Fujita and other renowned artists. Designed by architect Tadao Ando, the museum offers stunning exhibits and temporary exhibitions in a striking water and concrete building.

### Qwen2-VL Rerank: <Website 3>

### LLaVA-OneVision Brief Results:



**<Website 1>**  
**Title:** Akita Museum of Modern Art - Wikipedia  
**Snippet:** Akita Museum of Modern Art. Coordinates: 39u006017u203233u2033N 140u006032u203251u2033E. Akita Museum of Modern Art (u79cbu7530u770c u7acb u8fd1 u4ec3 u78e u8853 u9928, General information. Address: Tomigasawa-62-46, Akasaka, Town or city.



**<Website 5>**  
**Title:** Akita Museum of Modern Art - All You Need to Know BEFORE ... - Tripadvisor  
**Snippet:** Akita Museum of Modern Art. See all things to do. See all things to do. Akita Museum of Modern Art. 3.5. 19 reviews #12 of 55 things to do in Yokote. Art Museums. Write a review. Full view. All photos (39) Suggest edits to improve what we show. Improve this listing. Top ways to experience nearby attractions. Samurai Food Walk in Kakunodate. 2.



**<Website 2>**  
**Title:** MoMA  
**Snippet:** MoMA is a place that fuels creativity, ignites minds, and provides inspiration. Its extraordinary exhibitions and collection of modern and contemporary art are dedicated to helping you understand and enjoy the art of our time. Caption: The Museum of Modern Art Renovation and Expansion Designed by Diller Scofidio + Renfro in collaboration with Gensler.



**<Website 6>**  
**Title:** Iraq's Modern Art Collection, Waiting to Re-emerge  
**Snippet:** Such is the state of Iraq's modern art collection, renamed the National Museum of Modern Art in 2006 yet still an institution that exists mostly as an idea. That it exists at all is owed largely ...



**<Website 3>**  
**Title:** The Complete Guide to the Museum of Modern Art in New York City - TripSavvy  
**Snippet:** New York City's Museum of Modern Art, a.k.a. MoMA, was closed from June until October 2019 for a major \$450 million overhaul, both structurally and content-wise. Now expanded by 47,000 square feet (a 30-percent increase), MoMA boasts technological enhancements, including more than 200 digital audio guides you can listen to via free Wi-Fi, and an ever-changing temporary and permanent collection ...



**<Website 7>**  
**Title:** MoMA Visitor Tips and Advice - TripSavvy  
**Snippet:** The Museum of Modern Art, or MoMA for short, is one of New York City's most impressive museums. Designed by Japanese architect Yoshio Taniguchi, the building itself is as magnificent as its extensive art collection. The large six-story museum is home to many rotating exhibits and permanent art collections, so make sure to plan ahead to maximize your time.



**<Website 4>**  
**Title:** Locations, hours, and admission | MoMA  
**Snippet:** 10:30 a.m.-5:30 p.m. MoMA is open every day except Thanksgiving and Christmas. Exhibition and collection galleries close approximately 15 minutes prior to Museum closing. Our stores and restaurants have different hours.



**<Website 8>**  
**Title:** AD Classics: The Museum of Modern Art | ArchDaily  
**Snippet:** Share. The entrance to the Museum of Modern Art is tucked beneath a demure facade of granite and glass in Midtown Manhattan. Its clean, regular planes mark Yoshio Taniguchi's 2004 addition to the ...

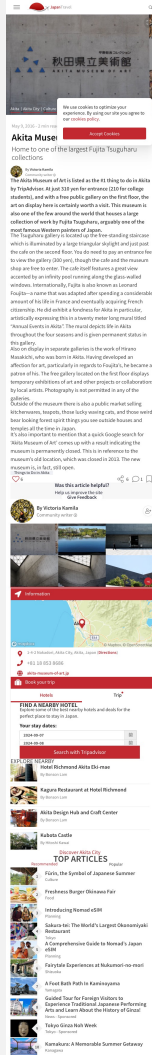
### LLaVA-OneVision Rerank: <Website 1>

Figure 10: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

## Round3 Summarization

### GPT-4o

#### Full-page Screenshot:



#### Content:

Akita Akita City Culture Akita Akita City Culture May 9, 2016 - 2 min read Akita Museum of Art Home to one of the largest Fujita Tsuguharu collections By Victoria Kamila Community writer The Akita Museum of Art is listed as the # 1 thing to do in Akita by TripAdvisor . At just 310 yen for entrance ( 210 for college students ) , and with a free public gallery on the first floor , the art on display here is certainly worth a visit . This museum is also one of the few around the world that Art ' comes up with a result indicating the museum is permanently closed . This is in reference to the museum ' s old location , which was closed in 2013 . The new museum is , in fact , still open . Things to Do in Akita 6 6 Share on Facebook Share on X ( Twitter ) Copy link to share 1 Was this article helpful ? Help us improve the site Give Feedback By Victoria Kamila Community writer Following Follow +6 Information 1-4-2 Nakadori , Akita City , Akita , Japan ( Directions ) +81 18 853 8686 akita-museum-of-art.jp Book your trip Hotels Trip Find a nearby hotel Explore some of the best nearby hotels and deals for the perfect place to stay in Japan . Your stay dates : Search with Tripadvisor Start your trip now When do you want to travel ? My dates are flexible Get started Explore nearby Hotel Richmond Akita Eki-mae By Bonson Lam Kagura Restaurant at Hotel Richmond By Bonson Lam Akita Design Hub and Craft Center By Bonson Lam Kubota Castle By Hitoshi Kawai Discover Akita City Top Articles Recommended Popular 1 Freshness Burger Okinawa Fair Food 2 Guided Tour Foujita ' s , he became a patron of his . The free gallery located on the first floor displays temporary exhibitions of art and other projects or collaborations by local artists . Photography is not permitted in any of the galleries . Outside of the museum there is also a public market selling kitchenwares , teapots , those lucky waving cats , and those weird bear looking forest spirit things you see outside houses and temples all the time in Japan . It ' s also important to mention that a quick Google search for ' Akita Museum of known as Leonard Foujita—a name that was adapted after spending a considerable amount of his life in France and eventually acquiring French citizenship . He did exhibit a fondness for Akita in particular , artistically expressing this in a twenty meter long mural titled “ Annual Events in Akita ” . The mural depicts life in Akita throughout the four seasons and is given permanent status in this gallery . Also on display in separate galleries is the work of Hirono Masakichi , who was born in Akita . Having developed an affection for art , particularly in regards to claim about Google saying it is closed . Wikipedia , Google maps , and the official website all say it is open ( in English and in Japanese ) . Perhaps there a few outdated websites that list it as closed . I used to go to the old one when I was in Akita 2011-2012 . The new one with poolside cafe looks nice . Reply Show all 0 replies 1 comment in total houses a large collection of work by Fujita Tsuguharu , arguably one of the most famous Western painters of Japan . The Tsuguharu gallery is located up the free-standing staircase which is illuminated by a large triangular skylight and just past the cafe on the second floor . You do need to pay an entrance fee to view the gallery ( 300 yen ) , though the cafe and the museum shop are free to enter . The cafe itself features a great view accented by an infinity pool running along the glass-walled windows . Internationally , Fujita is also for Foreign Visitors to Experience Traditional Japanese Performing Arts and Learn About the History of Ginza ! News - Sponsored 3 Introducing Nomad eSIM Planning 4 Sakura-tei : The World ' s Largest Okonomiyaki Restaurant Tokyo 5 Kamakura : A Memorable Summer Getaway Kanagawa 6 A Foot Bath Path In Kaminoyama Yamagata 7 Fairytale Experiences at Nukumori-no-mori Shizuoka 8 A Comprehensive Guide to Nomad ' s Japan eSIM Planning 9 Tokyo Ginza Noh Week Tokyo - Sponsored 10 Fürin , the Symbol of Japanese Summer Culture 1 A Guide to Japanese Visas Planning 2 Guide to Bringing Medicines Into Japan Planning

GPT-4o Summarize: 210 yen ✓

Figure 11: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

## Round3 Summarization

### Qwen2-VL

#### Full-page Screenshot:



#### Content:

Reviews US \$ 66.00 View Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte Address : 204-1 Aza Takuboshita Fuke Otsutsumi Distance : 445m Bar Pasaporte No reviews yet Other Cuisine View 2 . Kuidoraku Price : \$ 8.00 Address : 7-2 Ekimaecho, Yokote, Akita 013-0036 Distance : 2.28km Kuidoraku No reviews yet Bars/Bistros US \$ 8.00 View 3 . Korakuen Yokoteten Price : \$ 5.00 Address : It is 28-1, Sanmaibashi in Maego, Yokote-shi, Akita character Distance : 2.03km Korakuen Yokoteten No reviews yet Other Chinese Cuisine US \$ 5.00 View 4 . Ganso Kamiya Yakisoba Restaurant Address : Nakano-117-67 Oyashinmachi, Yokote, Akita 013-0051, Japan Distance : 1.17km Ganso Kamiya Yakisoba Restaurant No reviews yet Fast Food View Verified Reviews of Akita Museum of Modern Art 第二号爱人: 遇上秋田杆灯季, 还是挺热闹的一个节日在美术馆里面的话, 也有这种节日的气氛, 氛围都还是相当不错的, 而且的话里面虽然没有特别多的名画名品, 但是因为 是免费进入, 所以值得参观。Jedy Tan : 属于小众景点了, 秋田县本来就不大, 美术馆还在一个小 市里。但参观过能看出日本人近现代对西洋艺术的崇尚 xiaomoufa : 哈哈, 这是一个非常美丽的美术馆, 特别有意思的一个 景点。E30 \* \* \* 67 : 蛮大的美术馆, 里面画有些我们国家古代的味道, 蛮不错 的 Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium Price : \$ 14.30 Discount : \$ 2.04 Recommended sightseeing time : : 4-5 hours Address : Japan , 〒983-0013 and free . everything related to the ninjas were very fun . Edo Wonderland Nikko Edomura 4.5 / 5 43 Reviews No.3 of Best Things to Do in Nikko Theme Parks From US \$ 37.43 View Contents Duration Museum of Modern Art Opening Times Akita Museum of Modern Art Address Suggested Visit Duration for Akita Museum of Modern Art Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote 2 . Yokote Central Hotel 3 . Quad Inn Yokote 4 . Yokote Plaza Hotel Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte Hotels Flights Trains Cars Car Rentals Airport Transfers App Customer Support USD Search Bookings Sign in / Register Travel with Trip.com Travel Guide for Akita Museum of Modern Art in September ( Updated 2024 ) Akita Museum of Modern Art Opening Times Year round : 9:30-17:00 Akita Museum of Modern Art Address 62-46, Akasaka Tomigazawa | Inside Akita Furusato Mura , Yokote, Akita Prefecture Suggested Visit Duration for Akita Museum of Modern Art 1-2 hours Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote Address 2 . Kuidoraku 3 . Korakuen Yokoteten 4 . Ganso Kamiya Yakisoba Restaurant Verified Reviews of Akita Museum of Modern Art Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium 2 . Tsugaru-han Neputa mura Village 3 . Suntopia World 4 . Edo Wonderland Nikko Edomura Contents Akita Museum of Modern Art Opening Times Akita Museum of Modern Art Address Suggested Visit Duration for Akita Museum of Modern Art Featured Accommodation Near Akita Museum of Modern Art 1 . Hotel Plaza Annex Yokote 2 . Yokote Central Hotel 3 . Quad Inn Yokote 4 . Yokote Plaza Hotel Top Restaurant Picks Near Akita Museum of Modern Art 1 . Bar Pasaporte 2 . Kuidoraku 3 . Korakuen Yokoteten 4 . Ganso Kamiya Yakisoba Restaurant Verified Reviews of Akita Museum of Modern Art Also Popular With Visitors to Akita Museum of Modern Art 1 . Sendai Umino-Mori Aquarium 2 . Tsugaru-han Neputa mura Village 3 . Suntopia World 4 . Edo Wonderland Nikko Edomura Popular Travelogues Bangkok Travelogue | Manila Travelogue | Tokyo Travelogue | Taipei Travelogue | Hong Kong Travelogue | Seoul Travelogue | Kuala Lumpur Travelogue | Los Angeles Travelogue | Shanghai Travelogue \$ 14.30 View 2 . Tsugaru-han Neputa mura Village Price : \$ 3.38 Address : Japan , 〒036-8332 Aomori , Hirosaki , Kamenokomachi , 6 1 Distance : 0.63 mi Tsugaru-han Neputa mura Village No reviews yet From US \$ 3.38 View 3 . Suntopia World Price : \$ 8.85 Recommended sightseeing time : : 0.5-1 day Address : 1-1 Kubo , Agano , Niigata 959-2212 , Japan Distance : 5.06 mi What travelers say : M515shunyi1618 : Very suitable for family outings of playground , Ferris wheel , pirate ship are fun . Suntopia World 5 / 5 9 Reviews Amusement Ekimaecho Price : \$ 53.00 Distance : 2.23km Hotel Plaza Annex Yokote 4.3 / 5 37 Reviews -5 % US \$ 53.00 View 2 . Yokote Central Hotel Address : Heiwacho 9-10 Price : \$ 50.00 Distance : 2.91km Yokote Central Hotel 3.7 / 5 18 Reviews -5 % US \$ 50.00 View 3 . Quad Inn Yokote Address : Sekibata-52-1 Yasuda Price : \$ 58.00 Distance : 1.9km Quad Inn Yokote 4.1 / 5 14 Reviews US \$ 58.00 View 4 . Yokote Plaza Hotel Address : 7-1 Ekimaecho Price : \$ 66.00 Distance : 2.27km Yokote Plaza Hotel 3.9 / 5 10

Qwen2-VL Summarize: US \$ 66.00



Figure 12: Response and middle results comparison of GPT-4o (OpenAI, 2024b), Qwen2-VL-7B (Qwen Team, 2024), and LLaVA-OneVision-7B (Li et al., 2024a) in the end-to-end task.

