

LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models

Kaichen Zhang^{*,1,2} Bo Li^{*,1,2} Peiyuan Zhang^{*,1,2} Fanyi Pu^{*,1,2}
 Joshua Adrian Cahyono^{1,2} Kairui Hu^{1,2} Shuai Liu^{1,2} Yuanhan Zhang^{1,2}
 Jingkang Yang^{1,2} Chunyuan Li¹ Ziwei Liu^{1,2,✉}
¹LMMs-Lab Team ²S-Lab, NTU, Singapore
 {zhan0564, libo0013, peiyuan.zhang, fpu001, ziwei.liu}@ntu.edu.sg

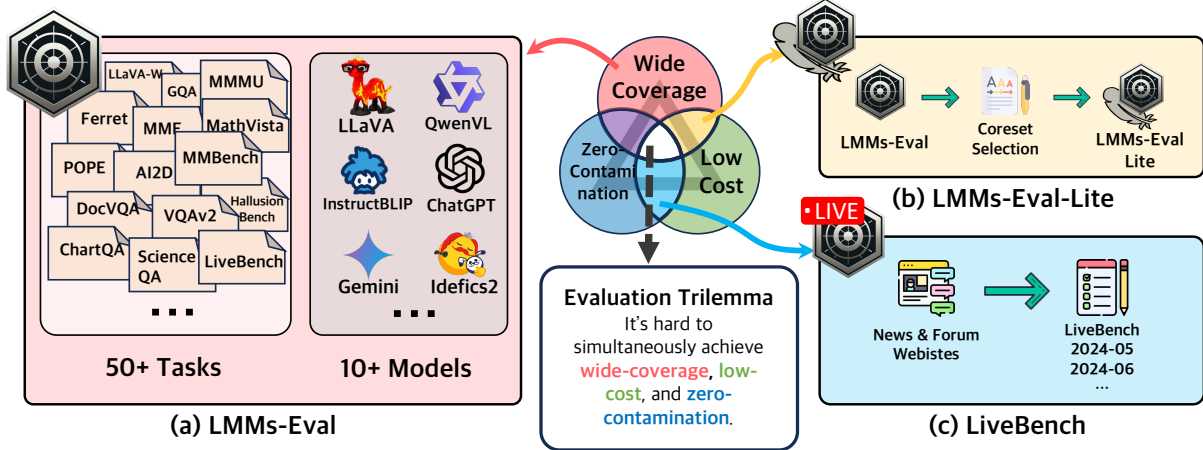


Figure 1: To best navigate the trilemma in LMM evaluation benchmarking, we contribute (1) LMMs-EVAL: a unified and standardized multimodal benchmark suite that encompasses over 50 tasks and more than 10 models, ensuring wide coverage; (2) LMMs-EVAL LITE: an efficient benchmark set with reliable and aligned results with the time-consuming full-set evaluation, addressing low-cost concerns; (3) LIVEBENCH: the evaluation benchmark with the latest information from news and forum websites, aiming to evaluate model’s zero-shot generalization ability on most recent events, thereby preventing contamination during evaluations.

Abstract

The advances of large foundation models necessitate wide-coverage, low-cost, and zero-contamination benchmarks. Despite continuous exploration of language model evaluations, comprehensive studies on the evaluation of Large Multi-modal Models (LMMs) remain limited. In this work, we introduce LMMs-EVAL, a unified and standardized multimodal benchmark framework with over 50 tasks and more than 10 models to promote transparent and reproducible evaluations. Although LMMs-EVAL offers comprehensive coverage, we find it still falls short in achieving low cost and zero contamination. To approach this evaluation trilemma, we further introduce LMMs-EVAL LITE, a pruned evaluation toolkit that emphasizes both coverage and efficiency. Additionally, we present Multi-modal LIVEBENCH that utilizes continuously updating news and online forums to assess models’ generalization abilities in the wild, featuring a low-cost and zero-contamination evaluation approach. In summary, our work highlights

the importance of considering the evaluation trilemma and provides practical solutions to navigate the trade-offs in evaluating large multimodal models, paving the way for more effective and reliable benchmarking of LMMs. We open-source our codebase and maintain leaderboard of LIVEBENCH at [Github](#) and [LiveBench](#).

1 Introduction

Good benchmarks guide AI development. Current large foundational models such as GPT-4 (OpenAI, 2024), Gemini (Gemini-Team, 2024), Claude (Anthropic, 2024), and many others (Team, 2024; Ormazabal et al., 2024; Mistral, 2024; Cohere, 2024) have demonstrated transformative capabilities, approaching or surpassing human-level performances in many tasks. In this context, benchmarks become both challenging and crucial to differentiate among the models and detect their weaknesses.

In the field of language models, exemplary works such as (Liang et al., 2022; Srivastava et al., 2022; Gao et al., 2023) aimed to comprehensively assess models across a wide range of dimensions.

*Equal contribution.

✉Corresponding author.

As generative AI evolves from language-centric to multimodal, a unified evaluation framework and a closer look at existing benchmarks are needed.

Transparent, standardized, and reproducible evaluations are crucial. We identify that there is so far no unified evaluation protocol in the field of LMM. Model publishers (Liu et al., 2023b; Team, 2024; Dai et al., 2023; Zhang et al., 2023; Li et al., 2023a) come up with custom evaluation pipelines, which often differ significantly in data preparation, output postprocessing, and metrics calculation, hindering transparency and reproducibility. To this end, we build a standardized and reliable benchmark suite to assess multimodal models in their entirety with LMMS-EVAL. LMMS-EVAL covers over 50 tasks in various scenarios to thoroughly assess more than 10 multimodal models with around 30 variants. It offers a standardized evaluation pipeline to ensure transparency and reproducibility. It also comes with a unified interface to facilitate the integration of new models and datasets.

Wide-coverage, low-cost, and zero-contamination benchmark is hard to achieve simultaneously. We believe it is an impossible triangle to evaluate models with wide coverage and low cost without making the benchmarks susceptible to contamination, as shown in Figure 1. For instance, the Hugging Face OpenLLM leaderboard (Team, 2023b) provides an economical way to evaluate language models across a wide range of tasks, but it is also prone to overfitting and contamination. The LMSys Chatbot Arena (Chiang et al., 2024) and AI2 WildVision (Lu et al., 2024b) offer robust and non-contaminated evaluation through real user interactions. However, it is expensive to gather tens of thousands of human preferences. In this work, we do not break this impossible triangle. Instead, we complement the evaluation landscape of LMMs by introducing LMMS-EVAL LITE and LIVEBENCH. By covering diverse sets of tasks and pruning unnecessary data instances, LMMS-EVAL LITE features a low-cost and wide-coverage LMM evaluation. On the other hand, LiveBench gathers the latest information from news and online forums to construct the test data, targeting an economical and generalizable way to do benchmarks.

In summary, we aim to offer a comprehensive view of the evaluations on multimodal models while presenting our observations and solutions. Our paper makes the following contributions:

(1) **LMMS-EVAL**: a unified multimodal models evaluation suite that covers over 50 tasks and more than 10 models with around 30 sub-variants. With LMMS-EVAL, we aim to streamline and standardize the evaluation process of multimodal models to ensure standardized comparisons between models.

(2) **LMMS-EVAL LITE**: an efficient evaluation set that provides reliable and aligned results with the time-consuming full-set evaluation. LMMS-EVAL LITE prunes unnecessary data instances to reduce the evaluation cost while maintaining the evaluation quality.

(3) **LIVEBENCH**: an evaluation benchmark that gathers the latest information from news and forum websites to evaluate models’ zero-shot generalization ability on the most recent events. LIVEBENCH aims to provide a low-cost and generalizable way to evaluate multimodal models.

2 LMMS-EVAL: A Unified Multimodal Models Evaluation Suite

Evaluation has often taken a significant amount of time in the model development cycle. In Section 2.1 we argue that existing evaluation pipelines in LMM contain much overhead and are not standardized. By introducing LMMS-EVAL, we reduce this overhead and scale up the evaluation. However, as we note in Section 2.2, there is still a trilemma in LMM evaluation that we cannot fully resolve but only find a better trade-off.

2.1 Scaling Evaluations with a Standardized Framework

Reducing the overhead Existing evaluations in LMMs are often done on a model-by-model and dataset-by-dataset basis (Liu et al., 2023b; Team, 2024). Researchers create custom inference scripts for their models across different benchmarks. While manageable for a single model and a few benchmarks, this process becomes highly inefficient when evaluating multiple checkpoints across ten or more datasets. Users need to manually launch each individual script to preprocess the datasets, inference models, and calculate final scores based on the outputs. Boilerplates are also abundant in the code. To address this, LMMS-EVAL follows the framework design of LM-EVAL-HARNESS (Gao et al., 2023) to allow for a one-command evaluation of multiple models and datasets. We preprocess and handle all the data needed during evaluation, ensuring a single data

Models	Parameters	AI2D	ChartQA	DocVQA	LLaVA ^W	Mathvista	MME	MMMU	RealworldQA
LLaVA-1.5-7B	7B	54.8	18.2	28.1	59.6	26.7	1859.0	35.3	55.8
LLaVA-NeXT-Vicuna-7B	7B	66.6	54.8	74.4	72.3	34.4	1841.8	35.1	57.8
LLaVA-NeXT-Mistral-7B	7B	60.8	38.8	72.2	71.7	37.4	1823.4	33.4	59.3
Qwen-VL-Chat	7B	45.9	60.1	66.3	21.2	24.6	1890.8	27.7	1.7
InstructBLIP-Vicuna-7B	7B	33.8	12.5	13.9	55.2	23.4	1508.7	28.4	37.4
LLaVA-NeXT-LLaMA3-8B	8B	71.6	69.5	78.2	80.1	37.5	1971.5	41.7	60.0
Xcomposer4K-HD	8B	78.1	80.6	90.8	74.2	57.3	2189.8	42.6	62.6
Idefics2-8B	8B	69.2	26.4	73.4	43.7	48.0	1792.1	39.7	25.5
LLaVA-1.5-13B	13B	59.5	18.2	30.3	66.1	26.4	1818.3	34.8	54.9
LLaVA-NeXT-Vicuna-13B	13B	70.0	62.2	77.5	72.3	35.1	1891.9	35.9	58.7
InstructBLIP-Vicuna-13B	13B	36.8	12.7	13.6	54.4	25.0	1529.6	33.7	42.4
InternVL-1.5	26B	79.0	83.8	92.4	90.2	61.5	2183.6	43.1	65.0
LLaVA-NeXT-34B	34B	74.9	68.7	84.0	88.8	46.0	2030.4	46.7	62.0
LLaVA-NeXT-72B	72B	77.4	77.0	84.4	89.2	46.6	2158.9	46.4	65.4
LLaVA-NeXT-110B	110B	80.4	79.7	85.7	90.4	49.0	2200.4	49.1	63.1
LLaVA-OV-0.5B	0.5B	57.1	61.4	73.7	74.2	34.8	1478.0	31.4	55.6
LLaVA-OV-0.5B(SI)	0.5B	54.2	61.0	75.0	71.2	34.6	1489.0	31.2	53.7
LLaVA-OV-7B	7B	81.4	80	90.2	90.7	63.2	1998.0	48.8	66.3
LLaVA-OV-7B(SI)	7B	81.6	78.8	89.3	86.9	56.1	2109.0	47.3	65.5
LLaVA-OV-72B	72B	85.6	83.7	93.1	93.5	67.5	2261.0	56.8	71.9
LLaVA-OV-72B(SI)	72B	85.1	84.9	93.5	93.7	66.5	2269.0	57.4	73.8

Table 1: An overview of selected results on LMMS-EVAL, achieved through a standardized and transparently reproducible pipeline.

source is used across different models for a standardized evaluation. Furthermore, detailed model outputs and results will be logged for future analysis.

Standardized evaluation Custom evaluation scripts also lead to another issue: the scores reported in different places are not directly comparable. For instance, (Li et al., 2023c) extracts model answers by comparing the output probabilities among the choices. It is counted correct so long as the ground-truth answer has the lowest perplexity among the choices (PPL-based). However, (Liu et al., 2023a) use the generation-based evaluation. An answer is counted as correct only if the model’s generation matches the option letter. To this end, we design a unified framework in LMMS-EVAL covering different evaluation setups. We believe there is no best setup but one needs to fix one when comparing results across different models. For a fair comparison, we also respect the chat template of the models if they are instruction-tuned. For reproducibility and transparency, a detailed log containing the evaluation setup, model generations, and score breakdown will be automatically logged. Since we designed a unified interface, new models and datasets can also be quickly added into LMMS-EVAL.

Equipped with these two core designs, we suc-

cessfully scaled up our evaluation to over 10 models and more than 50 datasets. We present partial results in Table 1 and the full supported models, datasets, and scores can be found in Appendix F and Appendix F.1. We believe that large-scale evaluations are crucial. They enable a comprehensive comparison across various aspects of model performance, revealing whether a model is a versatile performer or excels only in specific tasks. Additionally, large-scale, reproducible, and standardized evaluations are essential in ablation experiments to enhance our understanding of model architectures and training data.

2.2 The Evaluation Trilemma

Our ultimate goal is to find a wide-coverage, low-cost, and zero-contamination way to evaluate LMMS. However, even with LMMS-EVAL, we find it to be hard or even impossible. Specifically, once we scale the evaluation datasets to 50+, it becomes time-consuming to perform a full evaluation run on those datasets. Besides, those benchmarks are also susceptible to contamination during the training time (Yang et al., 2023a). As shown in Figure 1, we believe there is a trilemma in model evaluation. One can not achieve the three goals simultaneously but only find a trade-off. The LMSys Chatbot Arena (Chiang et al., 2024) and AI2 Wild-

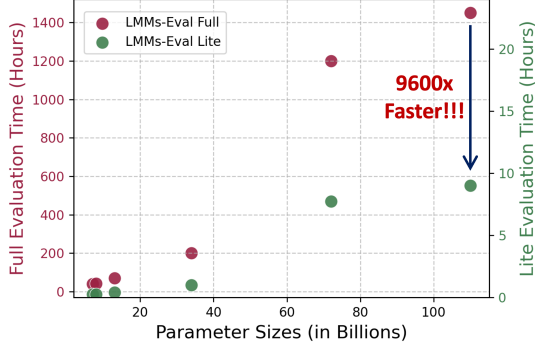


Figure 2: Evaluation cost demonstration on Full and Lite set.

Vision (Lu et al., 2024b) are foundational works in stressing wide coverage and anti-contamination. We present our solution to balance the other two sides of the triangle in Section 3 and Section 4.

3 LMMS-EVAL LITE: Affordable Evaluation with Broad Domain Coverage

We estimate the time to evaluate various LLaVA models on all LMMS-EVAL datasets in Figure 2. These evaluations were conducted using $8 \times A100$ GPUs with flash attention enabled. We replicate the model weights across GPUs and use data parallel by default. For models larger than 72B, we use pipeline parallelism (Huang et al., 2019) to load a single model across different GPUs.

We aim to construct a lite benchmark set that can provide useful and fast signals during the model development. If we can identify a subset of the benchmark where the absolute scores and relative rankings among models remain similar to the full set, we can consider it to be safe to prune the datasets. We thus present LMMS-EVAL LITE to complement the full datasets in LMMS-EVAL.

Lite set selection Let the benchmark be represented as $D = \{(x_i, y_i)\}_{i=1}^n$ and the scoring function underlying the benchmark system be denoted as S . Given a model f , let the response of the model to a particular question in the dataset be denoted as $f(x_i) = \hat{y}_i$. We aim to select a subset of the benchmark $V \in D$ such that

$$\min_{V: |V| \leq |D|} \left| \frac{1}{|D|} \sum_{i=1}^n S(y_i, \hat{y}_i) - \frac{1}{|V|} \sum_{i=1}^{|V|} S(y_i, \hat{y}_i) \right|$$

This objective function is equivalent to solving the k -Center problem (Sener and Savarese, 2018),

which seeks to identify a subset of data points that represent the full set. Thus, our problem is reformulated as finding representative points in x_i , which has been proven to be solvable as a k -Center problem (Sener and Savarese, 2018). Since solving the k -Center problem is NP-hard (Cook, 1997), we use a greedy algorithm to achieve a 2-OPT solution efficiently (details in Appendix D.4).

For k -center clustering, embeddings are extracted for each data point. While (Sener and Savarese, 2018) used CNN for image embeddings, we employed CLIP (Radford et al., 2021) for image embeddings and BGE-M3 (Chen et al., 2024a) for text embeddings, concatenating them to form the final embedding.

Dataset	Quire	k -means	Lite(Ours)
Flickr30k	0.97	0.79	0.91
AI2D	0.45	0.87	0.98
SeedBench	0.27	0.87	0.87
TextVQA	0.99	0.98	0.99

Table 2: Correlation results on multiple benchmarks and comparisons with k -means (Lloyd, 1982) and Quire (Huang et al., 2010)

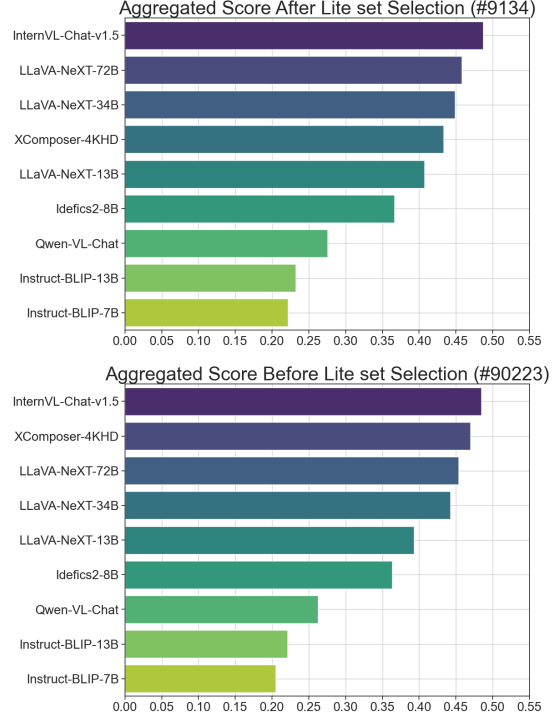


Figure 3: Results of LMMS-EVAL LITE across different models. The x -axis represent the weighted average percentage of scores that the model get across all the dataset.

To ensure our selected subset retains basic test-

ing abilities compared to the original benchmarks, we assess the correlation between the original scores and the lite set scores across six versions of LLaVA (Liu et al., 2023a). As shown in Table 2, our method maintain decent correlation results. Since the application of coreset selection in evaluating LMM datasets is limited and we are among the first to explore this approach to the best of our knowledge. There are only few methods for comparison. Additional results are provided in Appendix D.3.

Lite benchmark construction We refer to datasets from works like (OpenAI, 2023; Gemini-Team, 2024; Anthropic, 2024; Liu et al., 2023a) to construct LMMS-EVAL LITE, selecting 15 datasets across different task domains for broad coverage. To keep evaluation costs low, we apply a selection method to choose representative points from datasets with over 1500 data points. For MME (Fu et al., 2024), due to low correlation between the original and lite set scores, we retain the full version. In addition, we curate a new version of LMMS-EVAL LITE in Appendix D.3 that contains more datasets.

Score Aggregation To provide an overall signal to guide model development, we designed a strategy to aggregate the scores across different benchmarks in LMMS-EVAL LITE. Since different datasets and benchmarks come up with their own metrics, it is not reasonable to simply calculate the average score. Instead, we first normalize the scores from each dataset within a range of 100 and then calculate the average to be the final aggregated score. We report the aggregated score before and after the lite set pruning in Figure 3 to demonstrate the effectiveness of our selection method. Note that LMMS-EVAL LITE is not designed to fully compare the performance of different model families. Instead, it served as a tool to provide useful and low-cost signals during model training and ablations.

4 LIVEBENCH: From Static to Live Evaluation

4.1 Probing into Multimodal Data Contamination

LMMs are trained on massive amounts of data. For instance, Qwen-VL (Bai et al., 2023) leverages 1.4 billion pretraining data and CogVLM (Wang et al., 2024) uses 1.5 billion. However, research in both

LLMs (Zhang et al., 2024c; Wei et al., 2023) and LMMs (Chen et al., 2024b) has indicated that data contamination can significantly skew benchmark scores. This highlights the need for careful data management and validation to ensure accurate and fair evaluations.

We explore multimodal training within the LLaVA frameworks, utilizing two primary data types: (1) pretraining data to align visual and textual embeddings and train the vision encoder, and (2) high-quality, supervised finetuning data to improve diverse instruction-following capabilities. The re-annotation and conversion of large web and academic datasets into training materials frequently lead to issues of overlap and contamination. To address this, we developed an analytical tool to assess the overlap between training and benchmark data, showcasing our findings with data from (Liu et al., 2023a) with user data removed in it.

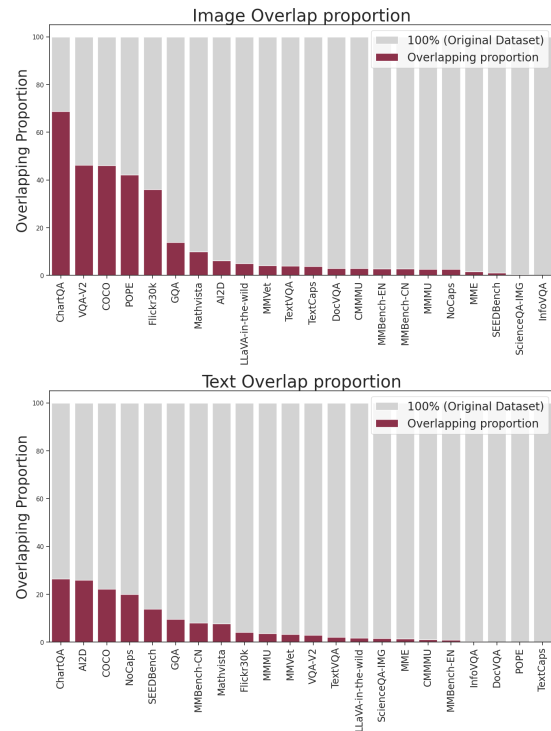


Figure 4: Contamination analysis in current evaluation benchmarks and LLaVA’s training data. Among the datasets with an overlap proportion exceeding 20%, including ChartQA, VQA-V2, COCO2014, and GQA, it has been confirmed that their training sets are included in LLaVA’s training data.

Text Overlap To measure text overlap, we use a string matching technique similar to those by GPT-4 (OpenAI, 2024), PaLM (Team, 2023a), and LLaMA (Touvron et al., 2023). Typically, an 8 ~ 13 n-grams range is used (Brown et al., 2020),

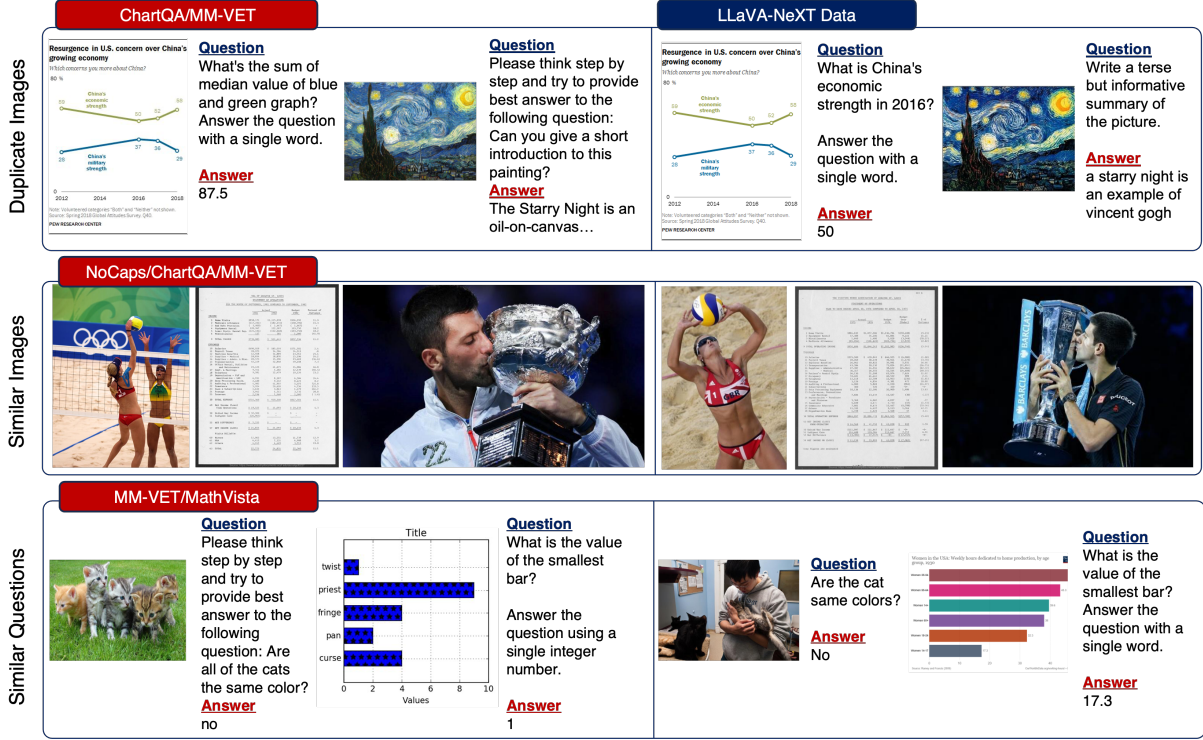


Figure 5: We present several cases of possible data overlapping in LLaVA-NeXT pretraining and supervised-finetuning data. We observed three types of data contamination (1) duplicate images (2) similar images (3) similar questions.

but we consistently use 8 n-grams for simplicity. We exclude any n-gram appearing more than 10 times in the training data, labeling these as *meaningless n-grams*. We also calculate an overlap ratio for each new n-gram candidate against our set of meaningless n-grams, excluding those exceeding a predefined threshold.

Image Overlap Contrary to text overlap, determining image overlap is a more challenging task. While it is common practice to compute image embeddings and then calculate their cosine similarity, selecting an appropriate threshold applicable to all datasets is difficult. Instead of computing similarity in the embedding space, we empirically find that using the pretrained SEED-tokenizer (Ge et al., 2023) leads to meaningful separation in detecting the overlap. We first tokenize each image into a 1-D sequence of 32 tokens. Similar to text, an 8-gram lookup table was constructed from those image tokens to detect image contamination. The occurrence of 8-gram overlap can be interpreted as approximately 1/4 of the image overlapping.

4.1.1 Results & Analysis on Decontamination

To evaluate the potential contamination of current benchmarks, we selected over 20 benchmarks, including AI2D (Kembhavi et al., 2016), ChartQA

(Masry et al., 2022), NoCaps (Agrawal et al., 2019), VQA v2 (Goyal et al., 2017), and LLaVA-in-the-wild (Liu et al., 2023b). We report the percentages of image and text overlap in Figure 4 for our selected datasets and more qualitative results in Figure 5. Our examination of both image and text overlaps has revealed three primary types of data contamination across various benchmarks.

Duplicate Images Instances of completely identical images between the training set and benchmark datasets were observed. This issue is exemplified by two identical images in ChartQA (Masry et al., 2022) and MM-Vet (Yu et al., 2023).

Similar Images Our image n-gram analysis has successfully identified the occurrence of visually similar images in both the training and benchmark datasets. Such similarities could lead to semantically similar questions, as demonstrated in examples from NoCaps (Agrawal et al., 2019), ChartQA (Masry et al., 2022) and MM-Vet (Yu et al., 2023).

Similar Questions We also observe recurring question structures in the training data that mirror those in the benchmark dataset. Although the corresponding images may differ, the similarity in

question structure could advantage the model in responding to benchmark queries.

4.2 Multimodal LiveBench

Traditional benchmarks rely on static evaluations with fixed questions and answers. While open-source models often outperform commercial ones like GPT-4V in benchmarks, they fall short in real user experience. Dynamic, user-oriented arenas like LMSys and WildVision are gaining popularity but face issues with prompt quality, difficulty, and noisy traffic, making consistent comparisons tough and costly. New benchmarks like Vibe-Eval (Padlewski et al., 2024) and LLaVA-Wilder (Li et al., 2024) use real-world data for more authentic testing, but as models continuously update from web data, there’s a risk of contamination in evaluation benchmarks.

We propose LIVEBENCH, a new evaluation framework that uses a dynamically updated dataset to prevent contamination and reduce costs. The evaluation data is collected from webpages, with an automated pipeline that gathers the latest global information from sources like news sites and community forums.

4.2.1 Dataset Curation Process

Data Collection From the Web To ensure the timeliness and authenticity of our information, we select sources from over 60 news outlets, including CNN, BBC, Japan’s Asahi Shimbun, and China’s Xinhua News Agency, as well as insights from forums like Reddit. A detailed list of these sources is provided in Appendix E.1.

Information Extraction The data collection pipeline is illustrated in Fig. 6, where the process begins by capturing screenshots of news website homepages. The information extraction consists of three main steps. 1) First, the model performs OCR to extract all text from the website. 2) The model is then instructed to identify significant images within the screenshot and extract relevant details about these images, such as the environment depicted, the actions and expressions of individuals, and the relationship between the images and the corresponding text. 3) Finally, the model is asked to specify what makes the information "newsworthy." For example, if the news is about the U.S. election, the model identifies what occurred in September 2024 that differentiates this news. Throughout the extraction process, we use Claude-3.5-Sonnet. All the prompts in this process can be found in Table 13

QA Generation The extracted information is then sent to the quiz model to generate questions and answers (QA). The model is prompted to create questions for four categories: (1) Concrete Recognition, (2) Real-world Application, (3) Analytical Understanding, and (4) Divergent Thinking & Creation. These categories are based on Bloom’s Taxonomy (Bloom et al., 1956). We prompt the model to produce challenging and innovative questions, along with criteria for scoring them. Detailed explanations of these categories and the prompts used to generate QA are provided in Table 9. An example QA with criteria can be found in Table 8

QA Checker & Finalizer To further curate high-quality QA pairs, we introduce the Checker and Finalizer models to refine the details of the QA pairs and validate the answers. The Checker model is mainly responsible for refining the questions and answers, restructuring them to ensure the questions are more answerable, verifiable, and challenging. It also ensures that the QA falls into the correct category. If the QA does not meet the requirements, the Checker model modifies the question and forwards it to the Finalizer. The Finalizer is mainly responsible for reformatting the question to enhance readability for human users. The prompt we use is included in Tables 11 and 12

QA Scorer The final part of our pipeline involves a scorer, which evaluates the QA pairs based on three criteria: Authenticity, Logical Coherence, and Clarity and Precision, assigning a score from 1 to 10. To balance data collection costs with evaluation efficiency, we collect approximately 500 questions each month and select 100 to 300 for the final LIVEBENCH problem set, based on those that exceed a certain score threshold. We also manually review the questions to remove any that are inappropriate. You can find the prompt in Table 10

We provide 4 examples for each category in Tables 21 to 24. It is important to note that the quality of our QA may still fall below that of human-curated answers, as we are aiming to build a dynamic evaluation pipeline that strikes a balance between cost and broad coverage.

4.2.2 Evaluation Metrics & Results on LIVEBENCH

We adopt the scoring criteria from LLaVA-Wilder (Li et al., 2024) and Vibe-Eval (Padlewski et al., 2024), using GPT-4o as the primary judge model. The judge assigns scores from 1 to 10 based

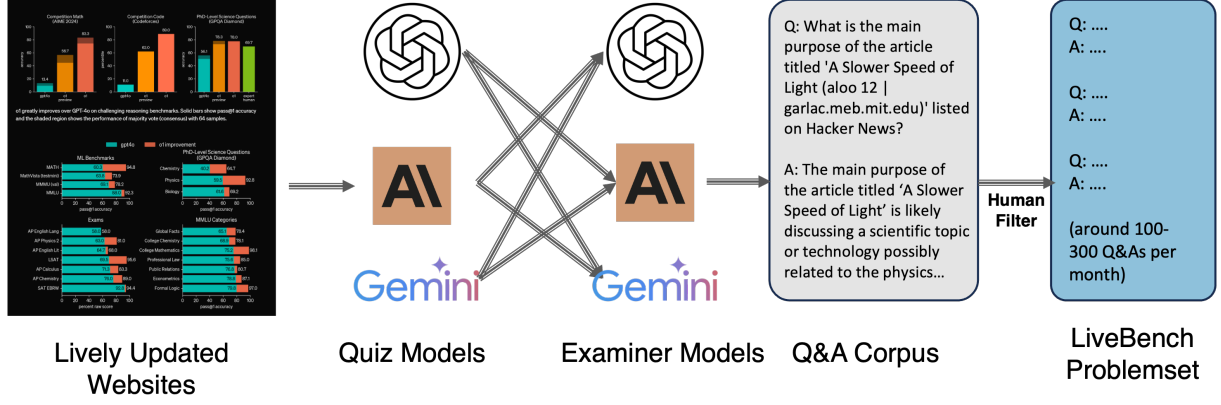


Figure 6: Overview pipeline for LIVEBENCH . We collect the latest information from the lively updated websites, organize the Q&A based on the information with the assistance of multimodal models, verify the Q&A with human annotators, evaluate the models with the Q&A corpus using different judge models, including human judges, and finally report the problemset.

on ground-truth answers and the scoring criteria. By leveraging established criteria, our evaluations are comprehensive and aligned with current standards. Detailed criteria and evaluation prompts are provided in Appendix E.3.

Model	Overall	Recognition	Analysis	Thinking	Realworld
LLaVA-1.5-7B	30.2	9.4	36.4	45.4	29.4
LLaVA-OV-0.5B	32.4	25.1	33.6	40.2	30.6
LLaVA-OV-7B	64.9	57.2	67.0	76.2	59.0
LLaVA-OV-7B-Chat	65.6	48.8	75.8	84.0	53.6
LLaMA-3.2-V-11B-Instruct	65.8	51.9	65.2	71.4	74.7
InternVL2-8B	69.6	65.6	74.8	77.5	60.4
LLaVA-OV-72B-Chat	75.0	62.0	87.8	83.8	66.6
Qwen2-VL-7B	79.2	74.2	82.8	87.4	75.2
Gemini-1.5-Flash	81.6	77.1	82.4	89.0	77.9
Gemini-1.5-Pro	84.5	85.4	83.8	88.6	80.1
Qwen2-VL-72B	85.9	86.7	88.8	89.0	79.2
Claude-3.5-sonnet	90.3	94.6	93.4	95.3	84.3
GPT4o-mini	91.9	94.6	93.4	95.3	84.3
GPT4o	92.0	91.7	93.8	94.8	87.6

Table 3: LiveBench-2024-09 Results.

The results in Table 3 indicate that the GPT-4 series models, including GPT-4o-mini and GPT-4o, are among the top performers, whereas the Gemini and Claude series models still outperform open-source models. GPT-4o has a large lead on recognition ability along with some small lead in other abilities. We provide a detailed case analysis in Appendix E.5 with many case studies to demonstrate how GPT-4o outperforms other models in many cases.

Open-sourced models are still far from achieving the level of GPT-4V. The *current superiority in benchmarks* can be attributed to the simplicity, fixed nature, or potential contamination of the evaluated scenarios (e.g., MME (Fu et al., 2024) and MMBench (Liu et al., 2024c)). These observations align with our hypothesis regarding the strengths

and limitations of commercial multimodal models like GPT-4V, which exhibit robust capabilities that existing benchmarks do not fully assess.

Specifically, our LIVEBENCH requires models to demonstrate strong zero-shot generalization abilities, as they must interpret continuously updated content from news and forum websites, highlighting the unique advantages of these commercial models.

While these findings may appear disadvantageous for competitors, they reveal the shortcomings of traditional benchmarks and emphasize the necessity for more comprehensive evaluations to accurately assess model performance. Benchmarking remains a crucial tool for driving progress in AI, and these results provide valuable insights for future contenders aiming to enhance their models.

5 Conclusions

In this work, we conducted a thorough reality check on the current evaluation pipeline and benchmarks for LMMs. We recognize the difficulties in the evaluation due to the *evaluation trilemma*. Although we cannot break this trilemma, we present three key contributions to find a better trade-off: 1) LMMs-EVAL, a unified evaluation suite for a standardized and large-scale LMM evaluation, 2) LMMs-EVAL LITE to balance low-cost evaluation with wide coverage, and 3) LIVEBENCH, a benchmark that transforms traditional static evaluation into a dynamic format to address potential data contamination in LMMs evaluation. We hope our LMMs-EVAL family makes a valuable contribution to the community towards the holistic

evaluation of LMMs.

6 Limitations

Through reality check, we explore the field of evaluation in LMMs and re-examine the evaluation process. Throughout our papers, we assume that the evaluation trilemma cannot be resolved. This suggests future work that goes deeper into finding a better trade-off among the sides of the trilemma or potentially overcoming it. Additionally, we address the issue of data contamination using a relatively simple method that requires access to the training data, while most research does not open-source their data. Future work may focus on methods that rely solely on the model and develop more efficient approaches.

Acknowledgments

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

- Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. no-caps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.
- Anthropic. 2024. [Introducing the next generation of claude](#). *Anthropic News*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşlılar. 2023. [Introducing our multimodal models](#).
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. [Scene text visual question answering](#). *Preprint*, arXiv:1905.13648.
- Benjamin S Bloom et al. 1956. Taxonomy of. *Educational Objectives*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023. [Benchlm: Benchmarking cross-style visual capability of large multimodal models](#). *Preprint*, arXiv:2312.02896.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2402.03216.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. [Are we on the right way for evaluating large vision-language models?](#) *Preprint*, arXiv:2403.20330.
- Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. [Websrc: A dataset for web-based structural reading comprehension](#). *Preprint*, arXiv:2101.09465.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#). *Preprint*, arXiv:2403.04132.
- Cohere. 2024. [Introducing command r+: A scalable llm built for business](#).
- W. Cook. 1997. *Combinatorial Optimization*. A Wiley-Interscience publication. Wiley.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). *Preprint*, arXiv:2305.06500.

- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.
- 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. 2024. *Mme: A comprehensive evaluation benchmark for multimodal large language models*. *Preprint*, arXiv:2306.13394.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. *A framework for few-shot language model evaluation*.
- Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. 2023. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*.
- Gemini-Team. 2024. *Gemini: A family of highly capable multimodal models*. *Preprint*, arXiv:2312.11805.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. *Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models*. *Preprint*, arXiv:2310.14566.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiad-bench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.
- Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.
- Sheng-jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. *Active learning by querying informative and representative examples*. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.
- Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. *Gpipe: Efficient training of giant neural networks using pipeline parallelism*. *Preprint*, arXiv:1811.06965.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. *ReferItGame: Referring to objects in photographs of natural scenes*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. *A diagram is worth a dozen images*. *Preprint*, arXiv:1603.07396.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. *Ocr-free document understanding transformer*. In *European Conference on Computer Vision (ECCV)*.
- Pang Wei Koh and Percy Liang. 2020. *Understanding black-box predictions via influence functions*. *Preprint*, arXiv:1703.04730.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. *What matters when building vision-language models?* *Preprint*, arXiv:2405.02246.
- Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. *Llava-next: Stronger llms supercharge multimodal capabilities in the wild*.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. *Otter: A multi-modal model with in-context instruction tuning*. *Preprint*, arXiv:2305.03726.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023b. *Seed-bench-2: Benchmarking multimodal large language models*. *Preprint*, arXiv:2311.17092.

- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023c. [Seed-bench: Benchmarking multimodal llms with generative comprehension](#). *Preprint*, arXiv:2307.16125.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023d. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023e. [Evaluating object hallucination in large vision-language models](#). *Preprint*, arXiv:2305.10355.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yanzhi Li, and Xiang Yue. 2024b. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. [Mmbench: Is your multi-modal model an all-around player?](#) *Preprint*, arXiv:2307.06281.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023d. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.
- Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.
- Frederic M Lord, Melvin R Novick, and Allan Birnbaum. 1968. *Statistical theories of mental test scores*. Addison-Wesley.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. [Learn to explain: Multimodal reasoning via thought chains for science question answering](#). In *Advances in Neural Information Processing Systems*.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2022b. [Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning](#). *Preprint*, arXiv:2110.13214.
- Yujie Lu, Dongfu Jiang, Wenhui Chen, William Wang, Yejin Choi, and Bill Yuchen Lin. 2024b. [Wildvision arena: Benchmarking multimodal llms in the wild](#).
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019a. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019b. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [Chartqa: A benchmark for question answering about charts with visual and logical reasoning](#). *Preprint*, arXiv:2203.10244.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706.

- Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. *corr abs/2007.00398* (2020). *arXiv preprint arXiv:2007.00398*.
- Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. [Coresets for data-efficient training of machine learning models](#). *Preprint*, arXiv:1906.01827.
- Mistral. 2024. [Mixtral 8x22b: Cheaper, better, faster, stronger](#).
- OpenAI. 2023. [Gpt-4v\(ision\) system card](#).
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Aitor Ormazabal, Che Zheng, Cyprien de Masson d’Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugénie Lamprecht, Hai Pham, Isaac Ong, et al. 2024. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*.
- Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*.
- Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. [Efficient benchmarking of language models](#). *Preprint*, arXiv:2308.11696.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. [tinybenchmarks: evaluating llms with fewer examples](#). *Preprint*, arXiv:2402.14992.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Ozan Sener and Silvio Savarese. 2018. [Active learning for convolutional neural networks: A core-set approach](#). In *International Conference on Learning Representations*.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. [Detecting pretraining data from large language models](#). *Preprint*, arXiv:2310.16789.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020a. [Textcaps: a dataset for image captioning with reading comprehension](#). *Preprint*, arXiv:2003.12462.
- Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020b. [Textcaps: a dataset for image captioning with reading comprehension](#).
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019a. [Towards vqa models that can read](#). *Preprint*, arXiv:1904.08920.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019b. [Towards vqa models that can read](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *arXiv preprint arXiv:2206.04615*.
- PaLM Team. 2023a. [Palm 2 technical report](#). *Preprint*, arXiv:2305.10403.
- Qwen Team. 2024. [Introducing qwen-vl](#).
- The HuggingFaceH4 Team. 2023b. [Open llm leaderboard - a hugging face space by huggingfaceh4](#).
- Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. [Hierarchical multimodal transformers for multi-page docvqa](#). *Preprint*, arXiv:2212.05935.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. [Anchor points: Benchmarking models with much fewer examples](#). *Preprint*, arXiv:2309.08638.
- Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. [Cogvlm: Visual expert for pretrained language models](#). *Preprint*, arXiv:2311.03079.
- Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. [Skywork: A more open bilingual foundation model](#). *Preprint*, arXiv:2310.19341.

- Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. 2023. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*.
- xAI. 2024. [Grok-1.5 vision preview](#).
- Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023a. [Rethinking benchmark and contamination for language models with rephrased samples](#). *Preprint*, arXiv:2311.04850.
- Yu Yang, Hao Kang, and Baharan Mirzasoleiman. 2023b. [Towards sustainable learning: Core-sets for data-efficient deep learning](#). *Preprint*, arXiv:2306.01244.
- Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014a. [From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions](#). *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014b. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. [Mm-vet: Evaluating large multimodal models for integrated capabilities](#). *Preprint*, arXiv:2308.02490.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2023. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). *Preprint*, arXiv:2311.16502.
- Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. [Mm-llms: Recent advances in multimodal large language models](#). *Preprint*, arXiv:2401.13601.
- Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhua Chen, and Jie Fu. 2024b. [Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark](#). *Preprint*, arXiv:2401.11944.
- Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024c. [A careful examination of large language model performance on grade school arithmetic](#). *Preprint*, arXiv:2405.00332.
- Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023. [Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition](#). *Preprint*, arXiv:2309.15112.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024d. [Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?](#) *Preprint*, arXiv:2403.14624.

A Related Work

Vision language benchmark Historically, benchmarks such as AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019a), TextCaps (Sidorov et al., 2020a), Flickr30k (Young et al., 2014a), and OK-VQA (Marino et al., 2019a) were used to assess computer vision model’s individual performance in captioning, optical character recognition, and visual question answering. With the emergence of Large Language Models (LLMs), Large Multimodal Models (LMMs) have been developed (Zhang et al., 2024a) to emphasize more comprehensive capabilities across vision and language. Subsequently, new benchmarks featuring increasingly challenging tasks and more holistic evaluation were proposed. For instance, benchmarks like ScienceQA (Lu et al., 2022a) and MathVista (Lu et al., 2024a) evaluate math and science abilities. SEED-Bench (Li et al., 2023c), CMMU (Zhang et al., 2024b), MMMU (Yue et al., 2023), and MM-Bench (Liu et al., 2024c), assess multiple heterogeneous dimensions of multimodal models/ In this paper, we aim to provide a comprehensive review of benchmarks from various fields.

Data contamination The issue of data contamination has emerged as a significant concern in the evaluation of Large Language Models (LLMs). Studies by (Yang et al., 2023a), (Wei et al., 2023), and (Zhang et al., 2024c) highlighted that data contamination poses a serious challenge for current LLMs and may lead to inaccuracies in accessing models’ real capabilities. Methods for data decontamination include assessing n-gram overlap (Brown et al., 2020), removing similar embedding points from datasets (Shi et al., 2024), or leveraging influential functions (Koh and Liang, 2020). However, the issue of data contamination in benchmarks for LMMs remains relatively unexplored.

Coreset benchmark With the development of numerous benchmarks, the demand for coreset versions across different benchmarks has become increasingly urgent. In LLM benchmarks, (Perlitz et al., 2024) employ stratified random sampling to select questions, while (Vivek et al., 2024) utilize the anchor points method for data point clustering. Other approaches, such as (Polo et al., 2024), utilize Item Response Theory (IRT) (Lord et al., 1968) to create embeddings for data points in benchmarks. In addition to these works, we have also inves-

tigated various active learning methods for efficiently and accurately constructing coresets. Quire (Huang et al., 2010) aims to select the most informative and representative points in the dataset, while (Mirzasoileiman et al., 2020), (Yang et al., 2023b), and (Sener and Savarese, 2018) focus on identifying coresets within the dataset.

B Broader Impacts

A comprehensive evaluation framework can help identify the limitations of existing multimodal models, preventing potential AI misuse. On the other hand, benchmarks can also introduce biases that may not reflect real-world scenarios. If the benchmarks are not representative of diverse applications and contexts, there is a risk that models optimized for these benchmarks may perform poorly in practical settings. Besides, automatic evaluations cannot replace expert human assessment in specialized fields such as medical imaging. The construction of LIVEBENCH uses real-world data crawled from the web. It could potentially lead to concerns regarding data privacy. The benchmarks we provide are meant for research purposes only and should be used with caution.

C Data Contamination

We present the details of the image overlapping in Table 4. Datasets such as ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2020), COCO (Lin et al., 2014), and VQAv2 (Goyal et al., 2017) were included in the LLaVA-NeXT (Liu et al., 2023a) training data and thus suffered the most from data contamination. Most of the benchmarks maintain a relatively low contamination proportion, with image and text overlap below 10%. POPE (Li et al., 2023e) was detected to have a high image overlapping ratio because it uses image sources from COCO (Lin et al., 2014).

C.1 More Qualitative Examples

We present more qualitative results here to demonstrate the data contamination problem in the dataset. We observe more identical images in benchmarks such as LLaVA^W (Liu et al., 2023b), MathVista (Lu et al., 2024a), and InfoVQA (Mathew et al., 2020). Similar images have also been another issue in different datasets; we present two more examples in NoCaps (Agrawal et al., 2019) and MM-Vet (Yu et al., 2023). Text overlapping can help us detect questions with similar sentence structure. Though

Table 4: Detailed image overlap and text overlap statistics accross different dataset

Dataset	Split	Image overlap (%)	Text overlap (%)
		LLaVA-NeXT Data	LLaVA-NeXT Data
Math & Science			
AI2D (Kembhavi et al., 2016)	test	6.09	25.97
MathVista (Lu et al., 2024a)	testmini	9.90	7.70
ScienceQA (Lu et al., 2022a)	img	0.35	1.54
Doc & Infographic			
ChartQA (Masry et al., 2022)	test	68.64	26.52
DocVQA (Mathew et al., 2020)	val	36.08	4.06
InfoVQA (Mathew et al., 2020)	test	0.14	0.39
Caption			
COCO2014 (Lin et al., 2014)	val	46.05	22.19
Flickr30k (Young et al., 2014a)	test	2.97	0.00
NoCaps (Agrawal et al., 2019)	val	2.53	19.98
TextCaps (Sidorov et al., 2020a)	val	3.79	0.00
VQA			
GQA (Hudson and Manning, 2019)	testdev-balanced	13.91	9.50
TextVQA (Singh et al., 2019a)	val	3.90	2.00
VQAv2 (Goyal et al., 2017)	val	46.21	2.90
Multi-task benchmark			
CMMMU (Zhang et al., 2024b)	val	2.89	1.11
MMBench (Liu et al., 2024c)	cn-dev	2.77	0.81
MMBench (Liu et al., 2024c)	en-dev	2.77	7.97
MME (Fu et al., 2024)	test	1.60	1.39
MMMU (Yue et al., 2023)	val	2.67	3.56
MMVet (Yu et al., 2023)	val	4.13	3.21
SEED-Bench (Li et al., 2023c)	all	1.11	13.84
Others			
LLaVA-W (Liu et al., 2023b)	test	5.00	1.67
POPE (Li et al., 2023e)	val	42.20	0.00

the images might not be similar enough, these similar questions might also be marked as in-domain questions. For example, we present two cases in MathVista (Lu et al., 2024a). Though not necessarily contamination or overlapping cases, the two images are both testing similar domain knowledge and may help the model to answer questions in the benchmarks.

D LMMS-EVAL LITE

D.1 Coreset Selection correlation

We compare the original scores and the selected dataset scores between the Lite version and the original datasets, calculating the correlation scores between them. We tried two different embeddings to perform k -center clustering. In addition to using CLIP (Radford et al., 2021) and BGE (Chen et al.,

2024a) embeddings, we also trained a LLaVA-Qwen 1.8B model following the training recipe of (Liu et al., 2023a) to embed image and text pairs simultaneously. For LLaVA embeddings, the last hidden states for all tokens were averaged into a single vector to serve as the feature vector for each data point. We report the correlation results for both embeddings in Table 5.

D.2 Dataset statistics in LMMS-EVAL LITE

We curated the first version of LMMS-EVAL LITE and present its correlation score and aggregation score in the paper. The exact plot of the correlation can be referred to Figure 8

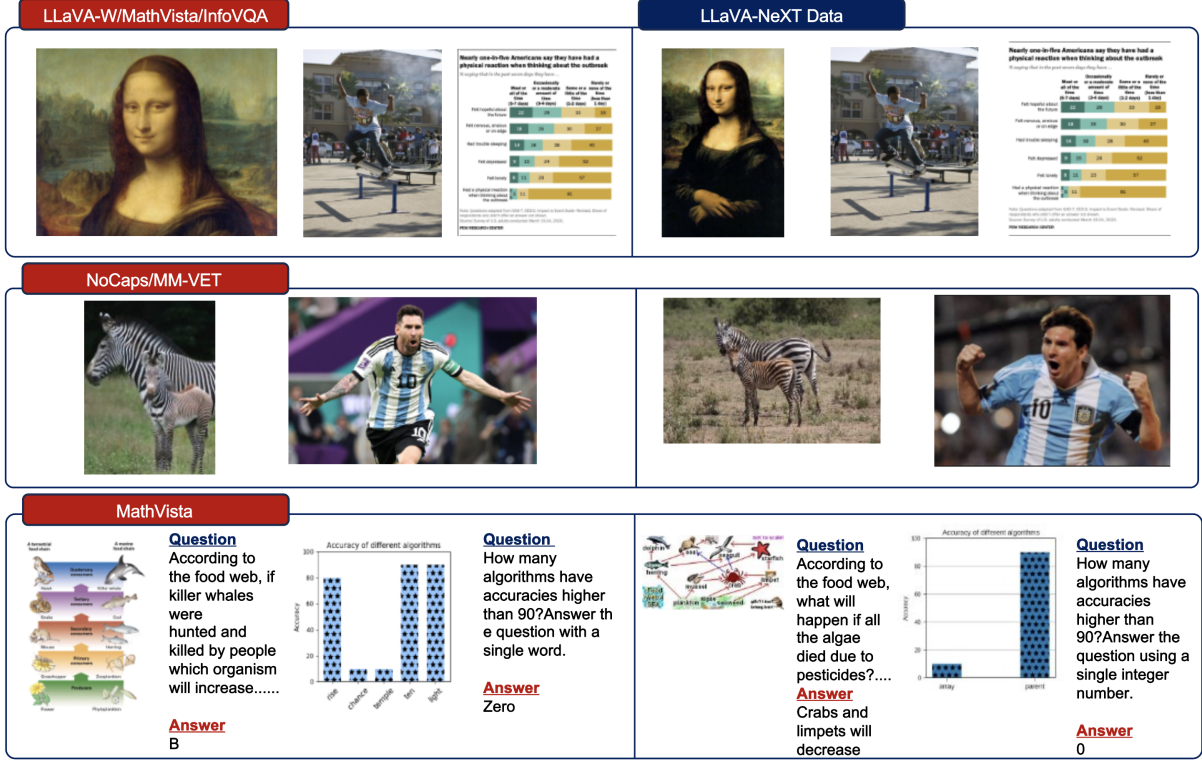


Figure 7: More qualitative results we found using our decontamination tools

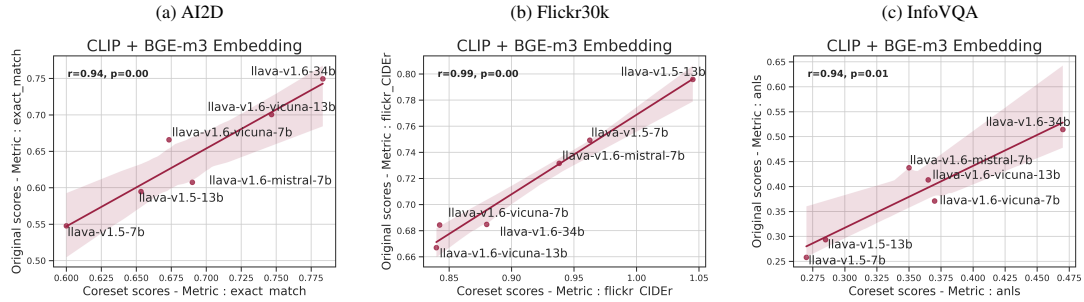


Figure 8: Correlation Graph between scores for our lite set and original scores

D.3 Curating more datasets in LMMS-EVAL LITE

We applied the same algorithm to additional datasets to develop a more comprehensive and diverse Lite version. In contrast to the original LMMS-EVAL LITE, our version incorporates more datasets, including COCO (Lin et al., 2014) and VQA (Goyal et al., 2017).

D.4 k-Center Greedy algorithm

The greedy algorithm we use for k -center clustering is detailed in Algorithm 1. In k -center clustering, the objective is to select k points among V vertices such that the maximum distance from any point in V to its nearest cluster center is minimized. In the employed greedy algorithm, a random point

is initially chosen as a center. Subsequently, the distance from this center to every other point is updated. The point with the maximum distance from the current centers is then selected and added to the center list. This process is repeated until k center points have been identified.

Algorithm 1 k -Center-Greedy

Input: data \mathbf{x}_i and $|V| = n$
Initialize $s = \phi$
while $|s| < n$ **do**
 $u = \arg \max_{i \in D \setminus s} \min_{j \in s} \Delta(\mathbf{x}_i, \mathbf{x}_j)$
 $s = s \cup \{u\}$
end while
return s

Table 5: The full correlation results we achieve using our selection methods

Dataset	Split	Lite Size	Original Size	Correlation	
				LLaVA Embedding	CLIP+BGE Embedding
Math & Science					
AI2D (Kembhavi et al., 2016)	test	300	3088	0.94	0.98
Doc & Infographic					
ChartQA (Masry et al., 2022)	test	400	2500	0.96	0.97
DocVQA (Mathew et al., 2020)	val	400	5349	0.99	0.99
InfoVQA (Mathew et al., 2020)	val	200	2801	0.94	0.94
Caption					
Flickr30k (Young et al., 2014a)	test	400	31784	0.99	0.91
NoCaps (Agrawal et al., 2019)	val	400	4500	0.99	0.98
TextCaps (Sidorov et al., 2020a)	val	300	3166	0.98	0.96
RefCOCO (Kazemzadeh et al., 2014)	val	500	8811	0.99	0.99
VQA					
TextVQA (Singh et al., 2019a)	val	300	5000	0.99	0.99
Multi-task benchmark					
SeedBench (Li et al., 2023c)	test	700	17990	0.77	0.87

E LiveBench Details

E.1 Website Candidates

To evaluate the performance and reliability of various news and information sources, a diverse set of websites has been selected for LIVEBENCH. We present the websites in Table 28. These websites span multiple categories, ensuring comprehensive coverage of different domains such as general news, business, technology, and international affairs. The list of candidate websites for LIVEBENCH includes prominent sources like BBC, CNN, Bloomberg, WSJ, and Reuters, among others. Each of these websites has been categorized based on its primary content focus. This categorization aids in the systematic evaluation of the content quality and the impact of imagery and reporting styles across different domains. It should be noted that this is a initial set of candidate websites and there may be changes depending on the situations of these websites.

E.2 Dataset Curation Prompts

This section outlines the dataset curation process, especially prompts used in different stages. First, the quiz model is provided with prompts to generate questions from raw website screenshots. The details of this prompt can be accessed at Table 9.

Once the candidate QAs are generated, we instruct the models to create corresponding scoring criteria for each question. The prompt used for this process is available at Table 10. Each question is

graded on a 10-point scale based on the provided criteria.

Lastly, we employ a checking model to verify the accuracy of the generated QAs. The prompt for this step is available at this Table 11.

E.3 Evaluation Prompts

We utilize GPT-4o as the default judge model due to its popularity and high-throughput API. Additionally, Claude-3.5-Sonnet and Gemini 1.5 Pro serve as alternative judge models. The final report results are scaled to an accuracy metric ranging from 0 to 100 based on the assigned scores.

Criteria are specified for each question, and we instruct the judge model to follow these criteria when determining the final score. An example of the criteria is provided at Table 8. Detailed judge prompts are available at Table 14.

E.4 Question Categorization and Examples

Building upon the principles of Bloom’s Taxonomy (Bloom et al., 1956), we aim to investigate the types of information that readers can extract from news content at different cognitive levels. Specifically, we focus on how readers interpret and process news reports, categorizing the information into the following hierarchical levels:

Concrete Recognition: At this level, the goal is to recognize facts and explain the fundamental concepts conveyed in the news. This may require models to possess optical character recognition (OCR)

Table 6: Overview of LMMS-EVAL LITE.

Task Domain	Dataset	Split	Full Size	Lite Size
Doc & Infographic Understanding	ChartQA	test	2500	400
	DocVQA	val	5349	400
	InfoVQA	val	2801	200
Image Understanding & Captioning	Flickr30k	val	31784	400
	NoCaps	val	4500	400
	TextCaps	val	3166	300
	RefCOCO	val	8811	500
Visual Question Answering	TextVQA	val	5000	300
Math & Science	MathVista	testmini	1000	1000
	AI2D	test	3088	300
Visual Dialogue	LLaVA-W	test	60	60
Multi-discipline	MME	cog. & percep.	2374	2374
	MMMU	val	900	900
	CMMMU	val	900	900
	Seed-Bench	test	17990	700
-	Total	-	90223	9134

capabilities to comprehend the context from provided screenshots and conclude the information. Example questions include: *What are the key points in this news story?* and *How would you explain the main event reported here?*

Realworld Application: At this level, individuals apply knowledge to real-world situations. Example questions include: *Please present this news in Arabic and output it in markdown format, Organize all the news on this page in the form of an HTML table, including the title, release time, and keywords, Sort out the exchange rate data and plot them using the Julia language, Please write a summary of the news in Vietnamese, and Can you give me an example of this update in Python?*

Analytical Understanding: This intermediate level emphasizes dissecting the news content to understand relationships and deeper meanings. Questions at this stage encourage analysis of the factors leading to an event and how it connects with other current issues. Example questions include: *What are the factors that led to this event?* and *How does this event relate to other current issues?*

Divergent Thinking & Creation: At the highest level, individuals engage in generating new ideas and synthesizing concepts to produce creative solutions. Questions at this level are designed to inspire divergent thinking and originality. Example questions include: *How could you create a new headline that captures the essence of the event differently?* and *If you were the reporter, how would you approach this story to provide a unique angle?*

We evaluate the model’s performance across these four progressively challenging levels, allowing us to assess its ability to transition from basic understanding to higher-order reasoning and creative thinking.

Specific examples corresponding to these levels are provided below. Tables 21, 22, 23, and 24 present representative examples within the LiveBench-2024-09 evaluation, illustrating the spectrum of cognitive demands posed by each level.

E.5 Case Analysis on LIVEBENCH

We present failure case analyses in Tables 15 and 18 to 20 to illustrate instances where current LMMs fail to respond accurately in our benchmark and the gap between these models and GPT-4o.

In Table 15, the model can not understand Japanese correctly and thus producing repeated nonsense sentences.

In Table 20, we see that the open-source model firstly made errors on identifying the correct numbers of the closing prices and then can not perform the arithmetic operations to get the average price, while GPT-4o manages to do so.

In Table 18, the model incorrectly matched the player names and their opponents. For instance, Karolina Muchova was supposed to play against Qinwen Zheng, but the model incorrectly stated that Muchova was leading against Anna Blinkova. Additionally, the model misidentified Qinwen Zheng as Qiang Wang, another Chinese tennis player. This demonstrates the model’s difficulty in recognizing small text on websites and its

Table 7: LMMS-EVAL LITE with more datasets

Task Domain	Dataset	Split	Full Size	Lite Size
Doc & Infographic Understanding	ChartQA	test	2500	500
	DocVQA	val	5349	500
	InfoVQA	val	2801	500
Image Understanding & Captioning	Flickr30k	val	31784	500
	NoCaps	val	4500	500
	TextCaps	val	3166	500
	RefCOCO	val	8811	500
	COCO	val	5000	500
Visual Question Answering	GQA	test	12578	500
	OKVQA	val	5046	500
	VizWiz-VQA	val	4319	500
	VQA-V2	val	214354	500
	TextVQA	val	5000	500
Math & Science	MathVista	testmini	1000	1000
	AI2D	test	3088	500
Visual Dialogue	LLaVA-W	test	60	60
Multi-discipline	MM-Bench	cn-dev	4329	500
	MM-Bench	en-dev	4377	500
	MME	cog. & percep.	2374	2374
	MMMU	val	900	900
	CMMMU	val	900	900
	Seed-Bench	test	17990	500
-	Total	-	340226	13734

tendency to hallucinate when failing to understand the image.

In Table 19, the model provided a detailed description but failed to summarize the main points. This indicates that the model may sometimes output unnecessary information and struggles with following instructions effectively.

F LMMS-EVAL Suite Information

Datasets on LMMS-Eval In previous research, benchmarks such as AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019a), TextCaps (Sidorov et al., 2020a), Flickr30k (Young et al., 2014a), and OK-VQA (Marino et al., 2019a) among many others, have been employed to assess a model’s performance in tasks such as captioning, optical character recognition (OCR), and visual QA. With the advent of Large Multimodal Models (LMMs), these have increasingly focused on broader capabilities spanning both vision and language, including reasoning (Lu et al., 2022a) and visual instruction following (Liu et al., 2023b). Consequently, new benchmarks featuring increasingly challenging tasks and more comprehensive evaluations have been proposed. For example, ScienceQA (Lu et al., 2022a) and MathVista (Lu et al., 2024a) assess mathematical and scientific competencies, while bench-

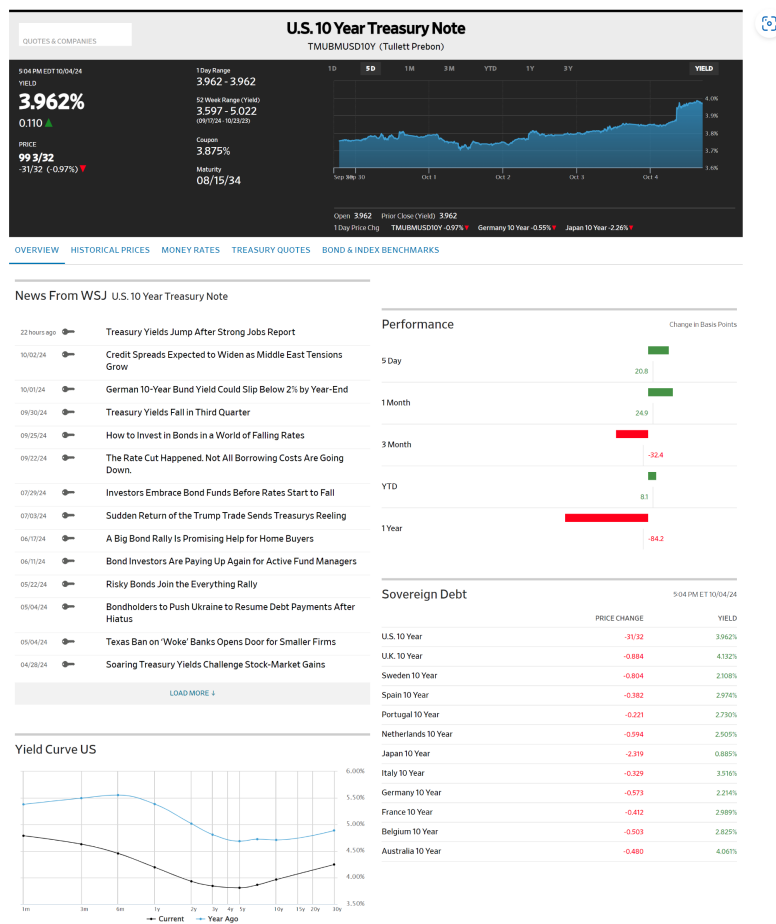
marks like SEED-Bench (Li et al., 2023c), CMMU (Zhang et al., 2024b), MMMU (Yue et al., 2023), and MM-Bench (Liu et al., 2024c) evaluate the multifaceted dimensions of multimodal models.

Models on LMMS-Eval To enable comparisons on new benchmarks for different models and to understand their capabilities across multiple tasks, we have supported over 10 models such as Fuyu (Bavishi et al., 2023), LLaVA (Liu et al., 2023b), Instruct-BLIP (Dai et al., 2023), InternVL (Chen et al., 2023), XComposer (Dong et al., 2024), Qwen-VL (Bai et al., 2023), MiniCPM (Hu et al., 2023), Idefics (Laurençon et al., 2024) and closed-source models such as GPT-4V (OpenAI, 2023), Gemini (Gemini-Team, 2024), Qwen-VL-Max (Team, 2024) and Claude (Anthropic, 2024).

F.1 Unified Evaluation Results with LMMS-EVAL

We present additional results using LMMS-EVAL here. Due to limited computational resources, we are only able to provide a holistic view of models from the LLaVA (Liu et al., 2023a) series. This demonstrates that achieving both wide coverage and low-cost evaluation simultaneously is not feasible, necessitating a balance between these two aspects.

Criteria Example



Question Identify the key details of the U.S. 10 Year Treasury Note provided in the image for October 4, 2024.

Ground Truth The key details of the U.S. 10 Year Treasury Note on October 4, 2024, include: - Yield: 3.962% - Price: 99 3/32, change of -31/32 (-0.97%) - Day Range (Yield): 3.962 - 3.962 - 52 Week Range (Yield): 3.597 - 5.022 - Coupon: 3.875% - Maturity Date: 08/15/34

Criteria The scoring criteria for the answer are as follows: 1. Yield: 2 points - 1 point for the correct yield value - 1 point for the correct formatting 2. Change in Yield: 2 points - 1 point for the correct change value - 1 point for the correct formatting 3. Price: 2 points - 1 point for the correct price value - 1 point for the correct formatting with the change 4. Day Range (Yield): 1 point - 1 point for the correct values 5. 52 Week Range (Yield): 1 point - 1 point for the correct values 6. Coupon: 1 point - 1 point for the correct value 7. Maturity Date: 1 point - 1 point for the correct value and formatting Deductions: - Minor mistakes in values: Deduct 0.5 points for up to 2 mistakes - Minor mistakes in formatting: Deduct 0.5 points for up to 2 mistakes

Table 8: An example of question, answer and criteria in LiveBench-09

Prompt : QA Generator

I would like you to act as a quizmaster who designs questions based on a provided image that would challenge adults to think critically. The image in question is a screenshot

- from the homepage or section of a news website. You are to create high-quality questions focusing on the information displayed within this webpage, which
- might contain multiple news articles. Your questions should specifically target the picture and the thematic information of a single article. Your question should
- be answerable, and checkable. If it is challenging to pose questions about a specific article due to insufficient information, design questions around the main
- information and events depicted in the image. Within your capabilities, try to make the questions more challenging.

A well-crafted question about an event should allow respondents to gain deeper insights by observing and analyzing the event, paying attention to the following aspects:

1. **Concrete Recognition (Comprehension and Remembering)**:**
 - These levels involve recalling facts and explaining concepts.
 - Example questions include:
 - "What are the key points in this news story?" (Remembering)
 - "How would you explain the main event reported here?" (Comprehension)
2. **Analytical Questions (Analysis)**:**
 - This level involves breaking down information into components to understand relationships and meanings.
 - Example questions:
 - "What are the factors that led to this event?"
 - "How does this event relate to other current issues?"
3. **Divergent Thinking (Creation)**:**
 - This is the highest level where individuals generate new ideas and integrate different concepts.
 - Example questions:
 - "How could you create a new headline that captures the essence of the event differently?"
 - "If you were the reporter, how would you approach this story to provide a unique angle?"
 - "Do you think the report's presentation was fair? Why or why not?"
4. **Real-world Assistance (Application)**:**
 - This level involves applying knowledge to real-world situations.
 - Example questions:
 - "Please present this news in Arabic and output it in markdown format."
 - "Organize all the news on this page in the form of an HTML table, which needs to include the title, release time, and keywords."
 - "Sort out the exchange rate data and plot them using Julia language."
 - "Please write a summary of the news in Vietnamese"
 - "Can you give me an example of this update in Python?" (Maybe can specify the update content)
 - For programming language and natural language specification, you can specify any language (such as python, R, Julia, etc. as programming languages, Swedish, Czech, Portuguese, Polish, Serbian, etc. as natural languages).

Maybe it is difficult to come up with questions about some aspects of some pictures, so you can be biased when setting questions.

Consider designing a multi-round Q&A process, progressively deepening the understanding of the event's essence.

Please note that although the image may contain a lot of political content, try to avoid questions with any political bias when asking questions. Your questions should focus on

- understanding and thinking about the image, not on political opinions.

You should try to be innovative, and you may propose some difficult questions, as well as multiple-choice questions, fill-in-the-blank questions, or even image-text

- matching questions, and sequencing questions. Within your capabilities, try to make the questions more challenging.

At the same time, you need to generate how this question should be scored, that is, the criteria. Each question is scored as \$0\sim 10\$, and the correct answers should be

- scored as \$10\$.
- Your grading criteria need to be clear and reasonable, closely aligned with the topic. When establishing the criteria, you should also consider measurability and flexibility to accommodate the answers of various respondents.

Table 9: The prompt that use to generate QA pairs

Prompt : Scorer

Based on the multi-round Q&A regarding the image, please evaluate each question and answer from the multi-round Q&A based on the image for their authenticity (whether

- the information can be directly obtained from the image or reasonably inferred) and logical coherence. For each Q&A pair, provide a rating from 1 to 10, where
- 1 indicates very poor and 10 indicates excellent. Additionally, please provide a brief explanation for each rating.

Here are the criteria for evaluating the Q&A pairs:

- ### 1. Authenticity (5 points)
- **5 Points**:: The information is directly observable in the image or can be reasonably inferred with strong evidence.
 - **3 Points**:: The information has a plausible connection to the image but requires assumptions that are not strongly supported by the image.
 - **1 Point**:: The information cannot be observed or reasonably inferred from the image; it seems unrelated or speculative.

- ### 2. Logical Coherence (3 points)
- **3 Points**:: The answer logically follows from the question and maintains consistency with the image context.
 - **2 Points**:: There are minor logical gaps or inconsistencies in the answer relative to the question.
 - **1 Point**:: The answer is logically inconsistent or contradictory to the question or image context.

- ### 3. Clarity and Precision (2 points)
- **2 Points**:: The question and answer are clearly articulated and precisely address specifics of the image.
 - **1 Point**:: The question or answer is somewhat vague or overly general, lacking specific details related to the image.
 - **0 Points**:: The question or answer is unclear or too ambiguous to determine relevance to the image.

Each Q&A pair can score a maximum of 10 points. The sum of points from these three categories determines the final score for each pair. Provide a brief explanation for each

- rating, focusing on how well the Q&A adheres to these criteria.

Table 10: The prompt that use to score the QA pairs

Prompt : Checker

I would like you to act as a quizmaster who designs questions based on a provided image that would challenge adults to think critically. The image in question is a screenshot

- from the homepage or section of news website. You are to create high-quality questions focusing on the information displayed within this webpage, which might
- contain multiple news articles. Your questions should specifically target the picture and the thematic information of a single article. Your question should be
- answerable, and checkable. Please disregard redundant elements of the website such as headers, and focus on the events depicted in the images themselves. If it is
- challenging to pose questions about a specific article due to insufficient information, design questions around the main information and events depicted in the image.

Now, you are given a screenshot of the homepage of a news website, with a already generated question and answer. Your task is to refine the question and answer, and

- refactor them to make the question more answerable, checkable, and challenging. If you don't think the question is good, please provide a new question and
- answer.

Note that the subtask must be one of these four:

- Concrete Recognition
- Analytical Questions
- Divergent Thinking
- Real-world Assistance

If you think the question does not correspond to the subtask, you have two options:

1. Modify the question to correspond to the subtask.
2. Modify the subtask to correspond to the question.

However, you should not change the original question's subtask unless the original subtask is not one of these five. If you feel the original question's subtask does not match

- the question, modify the question to match the subtask instead of rewriting the subtask.

Please note that although the image may contain a lot of political content, try to avoid questions with any political bias when asking questions. The question should focus on

- understanding and thinking about the image, not on political opinions. Within your capabilities, try to make the questions more challenging. However, you also
- need to consider the gradability of the questions you set. It is reiterated that what you need to assess is the ability to understand the news webpage, not politics.

You should try to be innovative, and you can also try different types of questions, like multiple-choice questions, fill-in-the-blank questions, or even image-text matching

- questions, and sequencing questions if possible. Within your capabilities, try to make the questions more challenging.

If you think the question is not good, or it is not answerable, please provide a new question and answer.

Reminder again that you cannot change the original subtask unless the original subtask is not one of the five listed above.

Table 11: The prompt that use to check the QA pairs

Prompt : Finalizer

You are a question setter, and your task is to finalize the question, answer, and scoring criteria. Make sure:

1. The criteria should be a natural language, don't use dict / json format for the criteria, human cannot understand it.
2. You can use bullet points / numbers to the list / yaml format to the criteria. But don't use python-like format.
3. If the answer is in dict format, but there is no need to answer in dict format (means there is a way to answer in natural language, the question do not specify to answer in dict format), you should convert it to natural language.
4. If the whole criteria is in other language, change it to English. But if you think some words should be in other language, you can keep it in that language. If question or answer is in other language, you don't need to change it.
5. The scoring criteria are rational and facilitate the accurate assessment of responses.
6. The full score for the scoring criteria must be 10 points, and it must directly relate to the specific answer.
7. The question is clear and unambiguous.
8. The answer is correct and reasonable (although the original ground truth answer is mostly correct, it may not be perfect, and sometimes the answer maybe incorrect).

Some tips:

1. For some extremely hard open-ended questions where answers may vary, hitting all points perfectly may not be realistic. In such cases, you can relax the criteria slightly.
 - For example, if there are five possible points in an answer, but answering three adequately could merit full points. An other option is to change the question to a
 - multiple-choice / multi-select question. But remember, it only applies to extremely hard open-ended questions which are impossible to answer perfectly.
2. For some questions, changing the format might be beneficial. You can consider transforming them into different types of questions such as essay, fill-in-the-blank, ranking
 - (e.g., based on time, importance, etc.), or matching questions to enhance the difficulty and rationality of the scoring criteria. But a very important point is that
 - DO NOT CHANGE the question to multiple-choice questions. If the original question is multiple-choice, you need to change it to another type of question (e.g.,
 - open-source, fill-in-the-blank, etc.).

Table 12: The prompt that use to finalize the QA pairs

Prompt : Information Extractor

These are the images of the website that we have captured. Please extract the text from the website.
You should extract the text from the website as detailed as possible.
Only output the text extracted from the website, do not include any other information.

This is a screenshot from a news website. Your task is to identify the meaningful images in this screenshot and extract relevant information about these images, such as the
→ environment depicted, the actions and expressions of the people, and the connection between these images and the corresponding text. You need to think deeply
→ about these images and provide as much detailed and useful information as possible. Of course, it is also possible that the website is mainly text-based, and in
→ this case, there might not be much information to extract from the images. In such instances, you can approach the task from a textual perspective, analyzing the
→ website's content. For example, what is the theme of the website? What kind of information is contained in the text? If the website requires thoughtful analysis,
→ feel free to engage in deeper reflection and provide your insights.

What makes this website different from other websites? What is special about its news? Since it is a news website, where is the 'new' aspect reflected? Do not provide a
→ generalized answer; you need to give detailed responses based on the specific content of each news article and the accompanying illustrations.

For example, if the news is about a software update, what conveniences will this update bring to people? How can people use these new features? Perhaps there are also some
→ drawbacks? You need to come up with your own questions worth pondering about the website and describe in as much detail as possible your understanding of
→ what is 'new' on the website.

Table 13: The prompt that use to extract information from website

Prompt : Information Extractor

[Question]

{ prompt }

[Assistant Response]

{ generation }

[Ground Truth Response]

{ reference }

[System]

Rate whether the assistant response correctly matches the ground truth, in regards to the image above.

The rating should be 0–10, where 0 is incorrect and 10 is correct.

Below is the specific criteria for rating:

{ criteria }

Your response should be in the JSON format:

```
```json
{
 "Explanation": "(your explanation)",
 "Rating": "(int)"
}
```

Table 14: The judge prompt that used in evaluation.

## Case Analysis : LLaVA-1.5-7B (Concrete Recognition)



Question Sort all the news headlines that appear in the image from newest to oldest. Provide the headlines and their respective timestamps.

LLaVA-1.5-7B 1. 2014/09/26 14:00:00 - 2014/09/27 00:00:00 2. 2014/09/26 13:00:00 - 2014/09/27 00:00:00 3. 2014/09/26 12:00:00 - 2014/09/27 00:00:00 4. 2014/09/26 11:00:00 - 2014/09/27 00:00:00 5. 2014/09/26 10:00:00 - 2014/09/27 00:00:00 6. 2014/09/26 09:00:00 - 2014/09/27 00:00:00 7. 2014/09/26 08:00:00 - 2014/09/27 00:00:00 8. 2014/09/26 07:00:00 - 2014/09/27 00:00:00 9. 2014/09/26 06:00:00 - 2014/09/27 00:00:00 10. 2014/09/26 05:00:00 - 2014/09/27 00:00:00 11. 2014/09/26 04:00:00 - 2014/09/27 00:00:00 12. 2014/09/26 03:00:00 - 2014/09/27 00:00:00 13. 2014/09/26 02:00:00 - 2014/09/27 00:00:00 14. 2014/09/26 01:00:00 - 2014/09/27 00:00:00 15. 2014/09/26 00:00:00 - 2014/09/27 00:00:00

GPT4o-Score 0

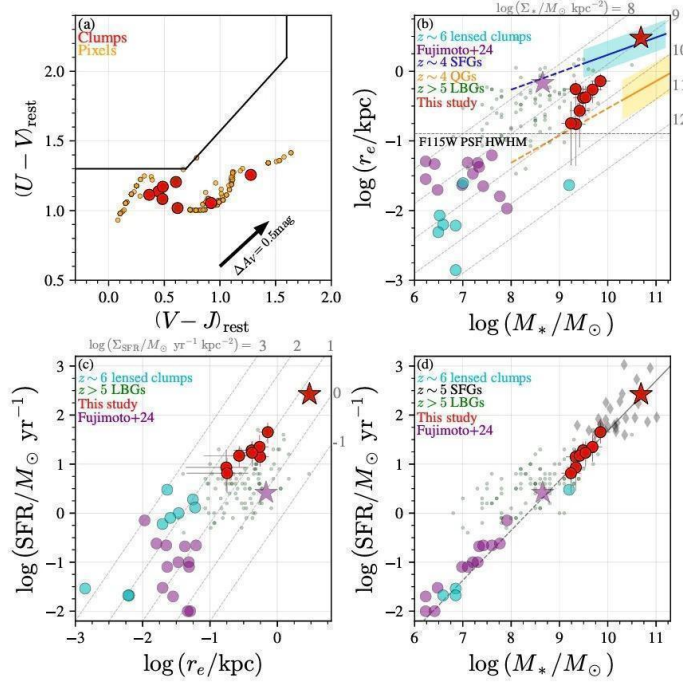
GPT4o-Rationale The assistant response did not match the ground truth at all. The assistant provided a list of timestamps that do not correspond to the actual timestamps in the image. Additionally, the headlines provided by the assistant are completely incorrect and do not match any of the headlines in the image. Therefore, the response is entirely incorrect.

GPT4o-Response Here are the news headlines sorted from newest to oldest along with their respective timestamps: 1. ひたちなかに里帰りしたロックン・出演者「一体感、半端じゃない」 - October 5, 17:00 2. 良い仕事は、不安のおかげ・テレビプロデューサー・佐久間宣行が著書 - October 5, 16:30 3. 35周年、「良い情報」科学の視点から「所さんの目がテン！」 - October 5, 16:30 4. (惜別) 真島茂樹さん・ダンサー・振付師 - October 5, 16:00 5. ガザの戦闘を止めるには・根本的解決策と「希望」、日本にできること - October 5, 16:00 6. (惜別) 田名網敬一さん・アーティスト - October 5, 16:00 7. 監督「私が見てもひどすぎる」・いま求められるドラマ「大幅カット」 - October 5, 12:00

Table 15: An example of the failure case of LLaVA-1.5-7B in LiveBench-09



## Case Analysis : Qwen-VL-72B-Instruct (Analytical Questions)

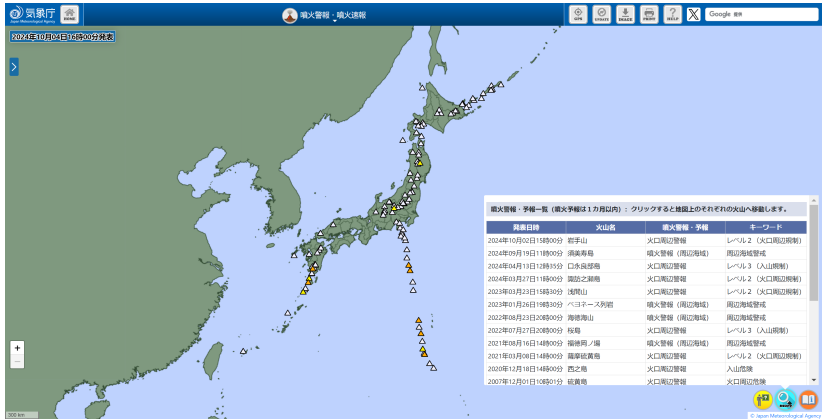


**Fig. 7.** Galaxy and clump characteristics. (a) Rest-frame  $UVJ$  color-color diagram. Orange and red dots indicate each pixel and clump, respectively. The black line indicates the threshold for classifying QG and SFG by Williams et al. (2009). The black arrow indicates the dust extinction corresponding to  $\Delta A_V = 0.5 \text{ mag}$ . (b) Size ( $r_e$ ) - mass ( $M_*$ ) relation. Red and purple dots represent each clump in CID-931 and  $z \sim 6$  clumpy system dubbed “Cosmic Grapes” in Fujimoto et al. (2024), respectively. Stars indicate the overall region of CID-931 and Cosmic Grapes (note that the red core component is not included in CID-931). Cyan dots indicate  $z \gtrsim 6$  lensed clumps (Welch et al. 2023; Vanzella et al. 2023; Fudamoto et al. 2024). Green dots represent high- $z$  Lyman break galaxies (LBGs, Morishita et al. 2024). The blue and orange lines and shaded regions indicate the size-mass relation for  $z \sim 4$  SFGs (Ward et al. 2024) and UVJ-selected QGs (Ito et al. 2024). The dashed lines indicate the extrapolation of each size-mass relation. The horizontal gray dashed line represents the half width at half maximum (HWHM) of the F115W PSF, while the diagonal gray dashed lines indicate fixed values of  $\Sigma_*$ . (c) Size ( $r_e$ ) - SFR relation. Each marker is the same as panel (b). The diagonal gray dotted lines indicate each  $\Sigma_{\text{SFR}}$ . (d) SFR -  $M_*$  relation. Each marker is the same as panel (b). The gray diamonds indicate  $z \sim 5$  SFGs (Khusanova et al. 2021; Xiao et al. 2023). The black solid and dashed lines represent the  $z \sim 5$  SFMS (Khusanova et al. 2021) and its extrapolation. The total  $M_*$  and SFR (red stars) are the sum of the values of each pixel. For the local  $r_e$ , we use  $r_e$  for the HSC  $i$ -band (section 2.2). Each clump in CID-931 is on the  $z \sim 5$  SFMS, massive with respect to the JWST SFG population, and more compact with higher  $\Sigma_*$  and  $\Sigma_{\text{SFR}}$  than typical high- $z$  SFGs.

Question	Based on graph (b) in the provided figure, what are the key differences between the galaxy characteristics of the ‘cosmic grapes’ clumps and the ‘CID-931 clumps’?
Qwen-VL-72B-Instruct	Graph (b) shows the size ( $r_e$ ) - mass ( $M_*$ ) relation for different galaxy populations. The ‘cosmic grapes’ clumps are represented by red dots, while the ‘CID-931 clumps’ are represented by purple dots. The key differences between the two populations are: 1. The ‘cosmic grapes’ clumps have a larger range of sizes ( $r_e$ ) compared to the ‘CID-931 clumps’. 2. The ‘cosmic grapes’ clumps have a larger range of masses ( $M_*$ ) compared to the ‘CID-931 clumps’. 3. The ‘cosmic grapes’ clumps are generally larger and more massive than the ‘CID-931 clumps’.
GPT4o-Score	1
GPT4o-Rationale	The assistant’s response does not correctly match the ground truth. The assistant incorrectly states that the ‘cosmic grapes’ clumps have a larger range of sizes and masses compared to the ‘CID-931 clumps’. Additionally, the assistant does not address the surface density differences between the two populations. The ground truth correctly identifies that ‘CID-931 clumps’ have higher stellar masses, larger sizes, and higher surface densities compared to ‘cosmic grapes’.

Table 16: An example of the failure case of Qwen-VL-72B-Instruct in LiveBench-09

## Case Analysis : Qwen-VL-72B-Instruct (Realworld Assistance)



### Question

Based on the volcanic warnings listed on the Japan Meteorological Agency's website, identify the volcano that was issued a Level 3 warning on April 13, 2024. Additionally, describe the type of warning and any precautions that were mentioned.

Qwen-VL-72B-Instruct

The volcano that was issued a Level 3 warning on April 13, 2024, is 水良部島 (Mizunoshima Island). The type of warning is "火山周辺警報" (Volcanic Surrounding Alert), and the precautions mentioned are "レベル 3 (入山規制)" (Level 3 (Mountain Entry Restriction)).

### GPT4o-Score

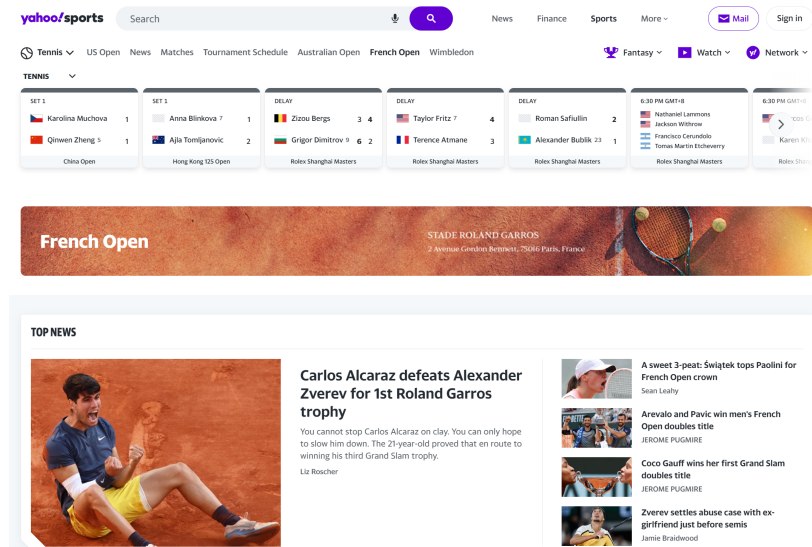
# O

## GPT4o-Rationale

The assistant incorrectly identified the volcano as 水良部島 (Mizunoshima Island) instead of 口永良部島 (Kuchinoerabujima), which results in 0 points for the first criterion. The type of warning was also incorrectly described as '火山周辺警報' (Volcanic Surrounding Alert) instead of '噴火周辺警報' (Eruption Warning - Surrounding Areas), resulting in 0 points for the second criterion. The assistant did not mention any of the specific precautionary measures (evacuation, restriction on entry, vigilance), resulting in 0 points for the third criterion.

Table 17: An example of the failure case of Qwen-VL-72B-Instruct in LiveBench-09

## Case Analysis : LLaVA-NeXT-OV-72B-Chat (Concrete Recognition)



Question

Analyze the ongoing tennis matches displayed on the webpage, detailing the players involved, their current scores, and the tournaments they are part of.

LLaVA-NeXT-OV-72B-Chat

The webpage displays several ongoing tennis matches from various tournaments. In the China Open, Karolina Muchova is leading against Anna Blinkova with a score of 1-0 in the first set. Meanwhile, in the Hong Kong 125K Open, Ajla Tomljanovic is trailing against Qiang Wang with a score of 0-1 in the first set. Over at the Rolex Shanghai Masters, Zizou Bergs is leading against Taylor Fritz with a score of 3-1 in the first set, and Grigor Dimitrov is leading against Alejandro Davidovich Fokina with a score of 2-0 in the first set. In the same tournament, Stefanos Tsitsipas is leading against Pablo Cuevas with a score of 2-0 in the first set, and Alexander Zverev is leading against Frances Tiafoe with a score of 2-0 in the first set. Lastly, in the Rolex Shanghai Masters, Rafael Nadal is leading against Jackson Withrow with a score of 2-0 in the first set, and Tomas Martincak is leading against Francisco Cerundolo with a score of 2-0 in the first set. These matches showcase a mix of established and up-and-coming players, highlighting the competitive nature of these tournaments.

GPT4o-Score

0

GPT4o-Rationale

The assistant response incorrectly identified the players, scores, and tournaments. The correct matches and scores are: Karolina Muchova vs. Qinwen Zheng (1-1), Anna Blinkova vs. Ajla Tomljanovic (7-2), Zizou Bergs vs. Grigor Dimitrov (3-4, delayed), Taylor Fritz vs. Terence Atmane (7-4, delayed), and Roman Safiullin vs. Alexander Bublik (2-1, delayed). The upcoming matches are Nathaniel Lammons/Jackson Withrow vs. Francisco Cerundolo/Tomas Martin Etcheverry and Marcos Giron vs. Karen Khachanov. The assistant response did not match any of these details correctly.

GPT4o-Response

Here are the details of the ongoing tennis matches displayed on the webpage: 1. **China Open** - **Karolina Muchova** vs. **Qinwen Zheng** - Current Score: 1-1 in Set 1 2. **Hong Kong 125 Open** - **Anna Blinkova** vs. **Ajla Tomljanovic** - Current Score: 7-2 in Set 1 3. **Rolex Shanghai Masters** - **Zizou Bergs** vs. **Grigor Dimitrov** - Current Score: 3-4 in Set 1 (Match is delayed) - **Taylor Fritz** vs. **Terence Atmane** - Current Score: 7-3 in Set 1 (Match is delayed) - **Roman Safiullin** vs. **Alexander Bublik** - Current Score: 2-1 in Set 1 (Match is delayed) Upcoming Matches: - **Nathaniel Lammons / Jackson Withrow** vs. **Francisco Cerundolo / Tomas Martin Etcheverry** - Scheduled for 6:30 PM GMT+8 at the Rolex Shanghai Masters - **Marcos Giron** vs. **Karen Khachanov** - Scheduled for 6:30 PM GMT+8 at the Rolex Shanghai Masters

Table 18: An example of the failure case of LLaVA-NeXT-OV-72B in LiveBench-09

## Case Analysis : LLaMA-3.2-Vision-11B-Instruct (Concrete Recognition)



Question

Summarize the main points of the news article regarding the golden owl treasure hunt in France. Be sure to include the duration of the hunt, the level of participation, the conclusion, and how the conclusion was announced.

LLaMA-3.2-Vision-11B-Instruct

The image shows a screenshot of an article from CNN Style, featuring a man in a white coat holding a golden owl statue. The headline reads "France's three-decade treasure hunt for golden owl has finally ended." \* The man is wearing a white coat and holding a golden owl statue. \* He is kneeling on the grass, with his left leg bent at a 90-degree angle. \* He is holding the owl statue in both hands, with his arms crossed over his chest. \* The background of the image is a sunset over a field, with trees and a cloudy sky. \* The sky is orange and yellow, with dark clouds in the distance. \* There are trees in the background, partially obscured by the man and the owl statue. \* The overall atmosphere of the image is one of excitement and celebration, as the man has finally found the treasure after a long search. \* The man's smile and the golden owl statue suggest that he is happy and proud of his discovery. \* The sunset in the background adds to the sense of triumph and accomplishment. The image suggests that the man has been searching for the golden owl for a long time and is finally able to hold it in his hands. The sunset in the background adds a sense of drama and excitement to the scene, emphasizing the significance of the discovery.

GPT4o-Score

0

GPT4o-Rationale

The assistant's response provides a detailed description of the image but fails to summarize the main points of the news article. It does not mention the duration of the hunt, the level of participation, the conclusion, or how the conclusion was announced. Therefore, it does not match the ground truth.

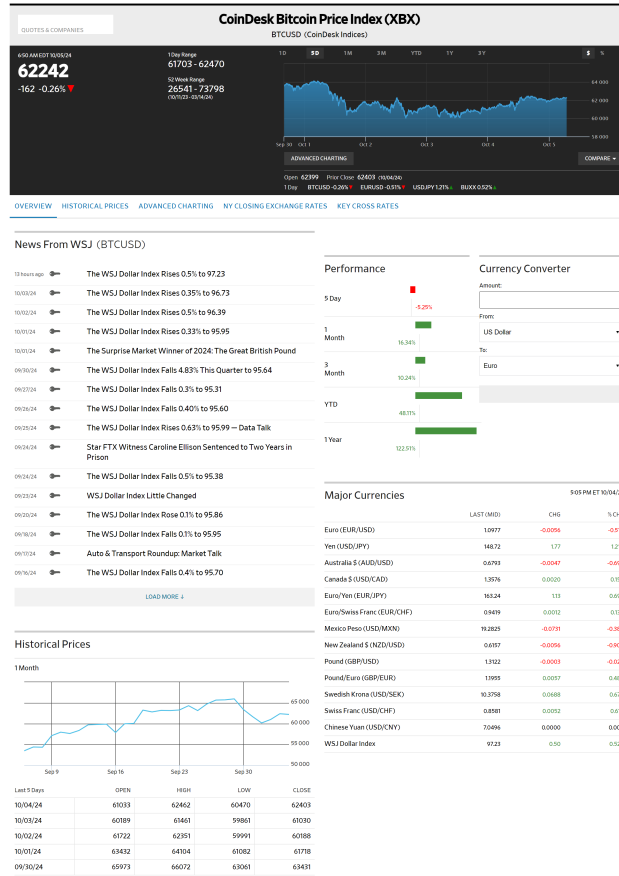
GPT4o-Response

The news article reports that France's golden owl treasure hunt, which lasted for 31 years and attracted thousands of enthusiasts across the country, has finally concluded. The end of the hunt was announced on Thursday through official social media accounts associated with the search, stating that the token required to claim the grand prize had been found.

Table 19: An example of the failure case of LLaMA-3.2-Vision-11B-Instruct in LiveBench-09



## Case Analysis : LLaVA-NeXT-OV-72B-Chat (Analytical Question)



Question

Based on the provided Bitcoin historical price data, calculate the average closing price for the last five days displayed.

LLaVA-NeXT-OV-72B-Chat

To calculate the average closing price for the last five days displayed, we add up the closing prices for each day and then divide by the number of days. The closing prices for the last five days are: 6243, 6103, 6018, 6178, and 62242. Adding these together gives us  $6243 + 6103 + 6018 + 6178 + 62242 = 77784$ . Dividing this sum by 5 gives us an average closing price of 15556.8 for the last five days displayed.

GPT4o-Score

0

GPT4o-Rationale

The assistant's response contains significant errors. It incorrectly lists the closing prices, including an erroneous value of 62242, which is not a closing price but the current price. The correct closing prices are 62403, 60470, 60188, 61718, and 63431. The assistant's summation and division are also incorrect, leading to a final average that is far from the correct value of \$61,642.

GPT4o-Response

To calculate the average closing price for the last five days displayed in the provided Bitcoin historical price data, we need to sum the closing prices of those five days and then divide by five. Here are the closing prices for the last five days displayed: 1. 10/04/24: 62403 2. 10/03/24: 61030 3. 10/02/24: 60188 4. 10/01/24: 61718 5. 09/30/24: 63431 Now, let's calculate the average:

$$\text{Average Closing Price} = \frac{62403 + 61030 + 60188 + 61718 + 63431}{5}$$

$$\text{Average Closing Price} = \frac{308770}{5} = 61754$$

So, the average closing price for the last five days displayed is 61,754.

Table 20: An example of the failure case of LLaVA-NeXT-OV-72B-Chat in LiveBench-09 for Analytical Question

## Question Type: Concrete Recognition

The screenshot shows the Yahoo Sports Tennis page. At the top, there's a navigation bar with 'yahoo/sports', a search bar, and links for News, Finance, Sports, and More. Below this, there's a 'Tennis' dropdown menu and a 'French Open' link. The main content area displays a grid of ongoing tennis matches with their current scores and tournament names. Below the matches, there's a banner for the 'French Open' at the 'STADE ROLAND GARROS'. Underneath the banner, there's a 'TOP NEWS' section featuring a large article about Carlos Alcaraz defeating Alexander Zverev for the 1st Roland Garros trophy, along with several smaller news snippets.

Tournament	Match	Score
China Open	Karolina Muchova vs. Qinwen Zheng	1-1 (Set 1)
Hong Kong 125 Open	Anna Blinkova vs. Ajla Tomljanovic	7-1 (Set 1)
Rolex Shanghai Masters	Zizou Bergs vs. Grigor Dimitrov	3-4 (Match delayed)
Rolex Shanghai Masters	Taylor Fritz vs. Terence Atmane	7-4 (Match delayed)
Rolex Shanghai Masters	Roman Safiullin vs. Alexander Bublik	2-1 (Match delayed)
Rolex Shanghai Masters	Nathaniel Lammons/Jackson Withrow vs. Francisco Cerundolo/Tomas Martin Etcheverry	6:30 PM GMT+8
Rolex Shanghai Masters	Marcos Giron vs. Karen Khachanov	6:30 PM GMT+8

**TOP NEWS**

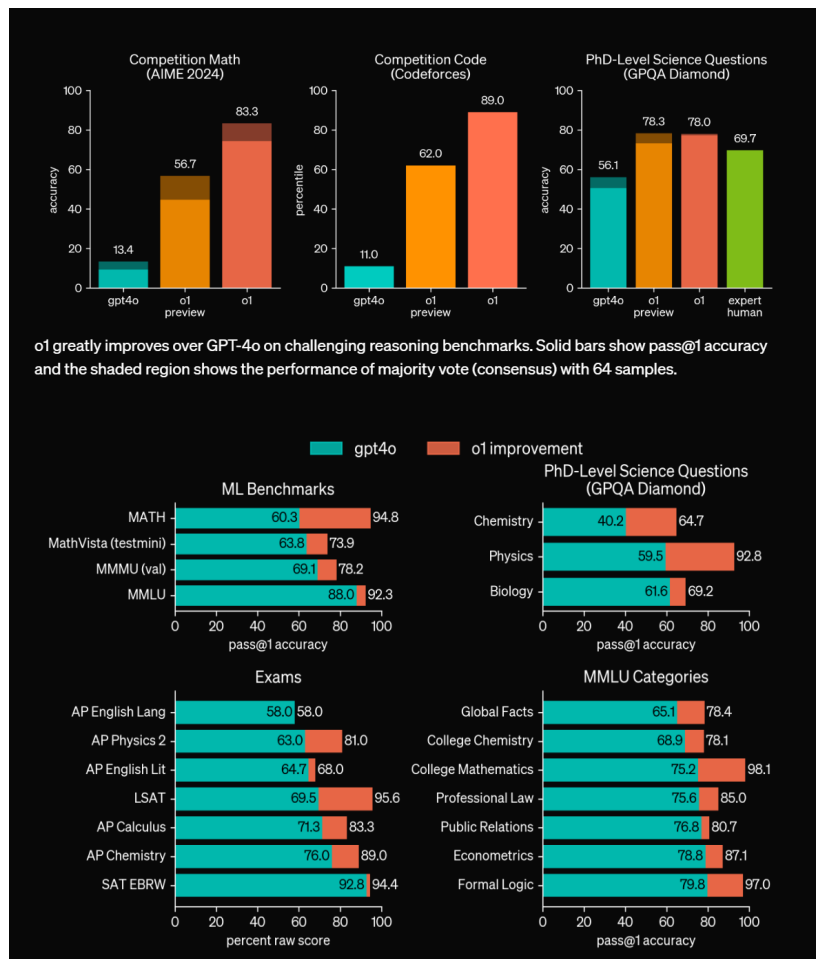
- Carlos Alcaraz defeats Alexander Zverev for 1st Roland Garros trophy**  
You cannot stop Carlos Alcaraz on clay. You can only hope to slow him down. The 21-year-old proved that en route to winning his third Grand Slam trophy.  
Liz Roscher
- A sweet 3-peat: Swiatek tops Paolini for French Open crown**  
Sean Leahy
- Arevalo and Pavic win men's French Open doubles title**  
JEROME PUGMIRE
- Coco Gauff wins her first Grand Slam doubles title**  
JEROME PUGMIRE
- Zverev settles abuse case with ex-girlfriend just before semis**  
Jennie Braddwood

**Question** Analyze the ongoing tennis matches displayed on the webpage, detailing the players involved, their current scores, and the tournaments they are part of.

**Ground Truth** Here are the ongoing tennis matches, scores, and tournaments: 1. China Open: - Karolina Muchova vs. Qinwen Zheng - Current score: 1-1 in the first set. 2. Hong Kong 125 Open: - Anna Blinkova vs. Ajla Tomljanovic - Current score: 7-1 in favor of Anna Blinkova in the first set. 3. Rolex Shanghai Masters: - Zizou Bergs vs. Grigor Dimitrov - Current score: 3-4, match delayed. 4. Rolex Shanghai Masters: - Taylor Fritz vs. Terence Atmane - Current score: 7-4, match delayed. 5. Rolex Shanghai Masters: - Roman Safiullin vs. Alexander Bublik - Current score: 2-1, match delayed. Upcoming matches at 6:30 PM GMT+8: - Nathaniel Lammons/Jackson Withrow vs. Francisco Cerundolo/Tomas Martin Etcheverry - Marcos Giron vs. Karen Khachanov

Table 21: An example of Concrete Recognition question in LiveBench-09

## Question Type: Real-World Application



Question

Create an HTML table summarizing the improvements in 'ML Benchmarks' shown in the image. The table should include the benchmark names, gpt4o scores, and o1 scores. Ensure proper HTML structure and formatting, and include basic styling for better readability.

Ground Truth

```
<!DOCTYPE html> <html lang="en"> <head> <meta charset="UTF-8"> <meta
name="viewport" content="width=device-width, initial-scale=1.0"> <title>ML Bench-
marks Improvements</title> <style> table width: 50%; border-collapse: collapse;
margin: 25px 0; font-size: 18px; text-align: left; th, td padding: 12px; border-
bottom: 1px solid #ddd; th background-color: #f2f2f2; </style> </head> <body>
<h2>ML Benchmarks Improvements</h2> <table> <thead> <tr> <th>Benchmark</th>
<th>gpt4o Score</th> <th>o1 Score</th> </tr> </thead> <tbody> <tr> <td>MATH</td>
<td>60.3</td> <td>94.8</td> </tr> <tr> <td>MathVista (testmini)</td> <td>63.8</td>
<td>73.9</td> </tr> <tr> <td>MMMU (val)</td> <td>69.1</td> <td>78.2</td> </tr>
<tr> <td>MMLU</td> <td>88.0</td> <td>92.3</td> </tr> </tbody> </table> </body>
</html>
```

Table 22: An example of Real World Application question in LiveBench-09

## Question Type: Analytical Understanding

### NEWS Latin America

#### Children among 70 killed in Haiti gang 'massacre'

Members of one of the country's armed groups, the Gran Grif gang, raided a small town forcing residents to run for their lives.

14 hrs ago | Latin America



#### UK commitment to Falklands 'unwavering' despite Chagos deal

Governor Alison Blake's remarks come as Argentina vows 'concrete action' to ensure the UK overseas territory is handed to Buenos Aires.

1 day ago | Africa



#### 'It kills everything' - Amazon's indigenous people hit by record forest fires in Brazil

More than 62,000 sq km of rainforest has burned in Brazil, this year alone.

2 days ago



#### Six migrants die after Mexico soldiers open fire

Officials say the incident happened after a military patrol saw a vehicle carrying 33 migrants travelling at speed.

2 days ago | Latin America



#### Mexico's first female president vows 'it's time for women'

At her inauguration, the first female president of Mexico promises to focus on improving health and education.

3 days ago | Latin America

#### ► US election: The view from Mexico

BBC Monitoring's Luis Fajardo looks at how the Mexican media has been reporting on US elections.

3 days ago | Latin America

#### The fierce battle over the 'Holy Grail' of shipwrecks

Colombia and a US firm are fighting over a shipwreck in the Caribbean Sea that could be worth \$18bn.

2 days ago | Business

Question

Analyze the scene depicted in the image associated with the Haiti gang 'massacre' article. Describe the environment, the actions of the people, and the emotions conveyed. How do these elements support the narrative described in the article?

Ground Truth

The image related to the Haiti gang 'massacre' article depicts a chaotic environment with frantic activity. People are seen running and some are on motorcycles, suggesting a hurried evacuation. The background shows buildings, trees, and power lines in what seems to be a small town setting. The overall mood is one of panic and urgency, as the townspeople appear distressed and desperate to escape, which aligns with the article's description of residents being forced to flee due to a gang raid by the Gran Grif gang. The image effectively conveys the fear and chaos precipitated by such violent incidents, underscoring the severe impact on the community.

Table 23: An example of Analytical Question in LiveBench-09



## Question Type: Divergent Thinking & Creation

### IT Unemployment at New High

The jobless rate in tech hit 6% in August, the seventh time in the last eight months that the IT jobless rate surpassed the national unemployment rate.



Source: Bureau of Labor Statistics analysis by Janco Associates

Joblessness for IT workers is at its worst since the dot-com bubble burst in the early 2000s, said Victor Janulaitis, Janco's chief executive. And this time the culprit is AI, he said, which is causing "seismic change" comparable to that seen when personal computers came into wide use.

Job reductions in IT remain concentrated in traditional technology roles like managing back-end corporate systems, Janulaitis said, [while AI](#) and cybersecurity are providing pockets of growth.

"The roles that are most in demand right now are roles where candidates have expertise in security, AI and the cloud," said Steve Watt, chief information officer of Hyland Software.

Within AI, businesses are looking for workers who can help them [prepare company data for use with AI models](#), as well as design tools to actually use AI, said Vikram Nafde, CIO of Webster Bank.

Question	Evaluate the potential impact of AI on the IT job market, considering both the negative and positive effects described in the provided image and text. Discuss the short-term and long-term impacts, and suggest strategies for IT professionals to adapt to these changes.
Ground Truth	<p>The potential impact of AI on the IT job market is multifaceted, with both negative and positive effects, as illustrated by the provided image and text. <b>Negative Effects:</b></p> <ol style="list-style-type: none"> <li><b>Job Reductions in Traditional Roles:</b> - The IT jobless rate rose to 6% in August, surpassing the national unemployment rate multiple times over the past eight months. This suggests a significant disruption in traditional IT jobs, such as managing back-end corporate systems, due to AI's capabilities in automating and optimizing these tasks.</li> <li><b>Seismic Changes:</b> - Victor Janulaitis, Janco's chief executive, compares the impact of AI to the seismic changes seen when personal computers came into wide use. This kind of systemic shift often results in skills becoming obsolete, leading to higher unemployment among professionals who cannot quickly adapt to new technological paradigms.</li> </ol> <p><b>Positive Effects:</b></p> <ol style="list-style-type: none"> <li><b>Growth in AI and Cybersecurity Roles:</b> - Despite the job reductions in some areas, AI and cybersecurity are cited as providing pockets of growth. This indicates that as AI takes over routine tasks, there is increased demand for developing, managing, and securing AI systems.</li> <li><b>High Demand for Expertise in Security, AI, and the Cloud:</b> - Steve Watt of Hyland Software mentions that roles requiring expertise in security, AI, and the cloud are most in-demand. This reflects a shift in the skill sets that are valuable in the job market towards more advanced and specialized IT roles.</li> <li><b>New Roles in AI Model Preparation and Tool Design:</b> - Vikram Nafde of Webster Bank highlights the need for workers to prepare company data for AI models and design tools to leverage AI. These new and emerging roles provide opportunities for IT professionals to transition into more specialized fields that AI technology creates. ... {Remaining Ground Truth answer}</li> </ol>

Table 24: An example of Creation Question in LiveBench-09

Table 25: Dataset Statistics in LMMS-EVAL. This table categorizes the initial set of tasks, detailing their task domains, ground-truth types, instance counts, and splits. We provide a comprehensive overview of the diverse datasets employed, which cover various task domains and evaluation metrics.

Datasets	Task Domains	Ground-Truth Types	Instances	Splits
AI2D (Kembhavi et al., 2016)	Science,Diagram	Muiti-Choice	3088	test
BenchLMM (Cai et al., 2023)	Cross Style Understanding	Short Answer / Muiti-Choice	102	test
ChartQA (Masry et al., 2022)	Chart	Short Answer	2500	test
CMMMU (Zhang et al., 2024b)	Multi-task,World Knowledge	Free-form / Muiti-Choice	900/11000	val/test
COCO 2014 Caption (Lin et al., 2014)	Captioning	Short Answer	40775 / 40504	test / val
COCO 2017 Caption (Lin et al., 2014)	Captioning	Short Answer	40670 / 5000	test / val
DocVQA (Mathew et al., 2020)	Document	Short Answer	5349	test
Ferret (You et al., 2023)	Referring or Grounding Actions	Free-form Answer	120	test
Flickr30k (Young et al., 2014b)	Visual Understanding	Captioning	31783	test
GQA (Hudson and Manning, 2019)	Real-World/Compositional QA	Short Answer	12578	test / dev
Hallusion-Bench (Guan et al., 2023)	Multimodal Image-Context Reasoning	Yes or No	951	image
IconQA (Lu et al., 2022b)	Abstract Diagrams	Muiti-Choice / Short Answer	21489 / 21488	test / val
InfoVQA (Mathew et al., 2022)	Infographics understanding	Extractive / Numerical	2801	val
LLaVA-COCO (Liu et al., 2023b)	Conversation, Reasoning	Free-form Answer	90	test
LLaVA-W (Liu et al., 2023b)	Conversation, Reasoning	Free-form Answer	60	test
LLaVA-Wilder (Liu et al., 2024a)	Conversation, Reasoning	Free-form Answer	210/1020	test
LiveBench (Ours)	Webpage Understanding / Lively Updated	Free-form	dynamic	test
MathVista (Lu et al., 2024a)	Mathematical Reasoning / Understanding	Free-form / Muiti-Choice	1000	testmini
MathVerse (Zhang et al., 2024d)	Mathematical Reasoning / Understanding	Free-form / Muiti-Choice	3940	testmini
MMBench (Liu et al., 2023c)	Reasoning / Perception	Muiti-Choice	6666 / 4329	test / dev
MME (Fu et al., 2024)	Perception, Cognition	Yes or No	2374	test
MMMU (Yue et al., 2023)	Multi-task, World Knowledge	Free-form / Muiti-Choice	10500 / 900	test / val
MM-Vet (Yu et al., 2023)	Multi-task	Free-form	218	test
Multilingual-LLaVA-W	Multi-lingual Conversation,Reasoning	Free-form Answer	60	test
MultiDocVQA (Tito et al., 2023)	Document	Short Answer	5019 / 5187	test / val
NoCaps (Agrawal et al., 2019)	Novel Object Captioning	Short Answer	4500	val
OCRBench (Liu et al., 2023d)	Text Recognition	Short Answer	1000	test
OKVQA (Marino et al., 2019b)	knowledge-based visual QA	Short Answer	5046	val
OlympiadBench (He et al., 2024)	Reasoning	Short Answer	2126 / 6351	test-en / test-cn
POPE (Li et al., 2023e)	Hallucination	Yes or No	9000	test
Q-Bench (Wu et al., 2023)	Image Quality Assessment	Short Answer / Muiti-Choice	2990	test
RealWorldQA (xAI, 2024)	Real world scenarios QA	Muiti-Choice	765	test
Refcoco (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	5000 / 1975 / 1810 / 8811	bbox-test / A / B / val
Refcoco (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	5000 / 1975 / 1810 / 8811	seg-test / A / B / val
Refcoco+ (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	1975 / 1798 / 3805,	bbox-testA / B / val
Refcoco+ (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	1975 / 1798 / 3805	seg-testA / B / val
Refcocog (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	5023 / 7573	bbox-testB / val,
Refcocog (Kazemzadeh et al., 2014; Mao et al., 2016)	Referring Expression	Short Answer	5023 / 7573	seg-test / val
ScienceQA (Lu et al., 2022a)	Science, World Knowledge, Reasoning	Muiti-Choice	4241	test
ScreenSPOT (Cheng et al., 2024)	GUI Understanding / Navigation	Short Answer / Coordinates	1272	test
SEED-Bench (Li et al., 2023d)	Spatial and Temporal Understanding	Muiti-Choice	17990	test
SEED-Bench-2 (Li et al., 2023b)	Multi-disciplinary Knowledge	Muiti-Choice	24371	test
ST-VQA (Biten et al., 2019)	Highlevel Semantic Information Understanding	Short Answer	4070	test
SynthDoG (Kim et al., 2022)	Text Understanding	Free-form	500 / 500	val-en / val-zh
TextCaps (Sidorov et al., 2020b)	Text Understanding	Captioning	21953 / 3166 / 3289	train / val / test
TextVQA (Singh et al., 2019b)	Text Understanding	Short Answer	5000 / 5734	val / test
VisualWebBench (Liu et al., 2024b)	Webpage Understanding / OCR / Reasoning	Short Answer / Muiti-Choice	1536	test
VizwizVQA (Gurari et al., 2018)	Low Quality Image Understanding	Short Answer	8000 / 4319	test / val
VQAv2 (Goyal et al., 2017)	Visual QA	Free-form	447793 / 214354	test / val
WebSRC (Chen et al., 2021)	Structure of Webpage	Short Answer / Yes or No	40357 / 52826	test / dev

Table 26: Detailed Statistics of the Initial Set of Models in LMMS-EVAL. The models are categorized by their model family, with their inference parameters, model types (indicating whether they are open-sourced or accessed via API), and parallel types, which denote the strategy leveraged during the model inference.

Model Family	Model Version	Parameters	Model Type	Parallel Type
InstructBLIP	InstructBLIP-Vicuna-7B	7B	Open-sourced	Data
	InstructBLIP-Vicuna-13B	13B	Open-sourced	Data
Fuyu	Fuyu-8B	8B	Open-sourced	Data
Idefics	Idefics-2-8B	8B	Open-sourced	Data
MiniCPM	MiniCPM-V 2.8B	2.8B	Open-sourced	Data
XComposer	XComposer-4KHD	8B	Open-sourced	Data
InternVL	InternVL-1.5	26B	Open-sourced	Data
LLaVA	LLaVA-1.5-7B	7B	Open-sourced	Data
	LLaVA-1.5-13B	13B	Open-sourced	Data
	LLaVA-NeXT-Vicuna-7B	7B	Open-sourced	Data
	LLaVA-NeXT-Vicuna-13B	13B	Open-sourced	Data
	LLaVA-NeXT-Mistral-7B	7B	Open-sourced	Data
	LLaVA-NeXT-Yi-34B	34B	Open-sourced	Data
	LLaVA-NeXT-LLaMA-3-8B	8B	Open-sourced	Data
	LLaVA-NeXT-Qwen-72B	72B	Open-sourced	Model
	LLaVA-NeXT-Qwen-110B	110B	Open-sourced	Model
Qwen-VL	Qwen-VL-Chat-7B	7B	Open-sourced	Data
	Qwen-VL-Plus	N/A	Close-sourced, API	Data
	Qwen-VL-MAX	N/A	Close-sourced, API	Data
Gemini	Gemini-1.0-Pro	N/A	Close-sourced, API	Data
	Gemini-1.5-Flash	N/A	Close-sourced, API	Data
	Gemini-1.5-Pro	N/A	Close-sourced, API	Data
GPT4	GPT-4V	N/A	Close-sourced, API	Data
	GPT-4O	N/A	Close-sourced, API	Data
Claude	Claude-3-Haku	N/A	Close-sourced, API	Data
	Claude-3-Sonnet	N/A	Close-sourced, API	Data
	Claude-3-Opus	N/A	Close-sourced, API	Data

Table 27: More results using LMMS-EVAL

	Split	Metric	#Num	LLaVA-1.5-7B	LLaVA-1.5-13B	LLaVA-NeXT-mistral-7B	LLaVA-NeXT-vicuna-7B	LLaVA-NeXT-13B	LLaVA-NeXT-34B
COCO-Cap	cococap_val_2014	CIDEr	40,504	108.66	113.88	107.66	96.98	99.45	103.16
COCO-Cap	cococap_val_2017	CIDEr	5,000	110.38	115.61	109.22	99.93	101.99	105.89
DocVQA	val	ANLS	5,349	28.08	30.29	72.16	74.35	77.45	83.98
GQA	testdev_balanced_instructions	Acc	12,578	61.97	63.24	54.98	64.23	65.36	67.08
MultidocVQA	val	Anls/acc	5,187	16.65/7.21	18.25/8.02	41.4/27.89	44.42/31.32	46.28/32.56	50.16/34.93
NoCaps	nocaps_eval	CIDEr	4,500	105.54	109.28	96.14	88.29	88.27	91.94
OKVQA	val	Acc	5,046	53.44	58.22	54.77	44.25	46.27	46.84
POPE	test	F1 Score	9,000	85.87	85.92	86.79	86.4	86.26	87.77
ScienceQA	scienceqa-full	Acc.	4,114	70.41	74.96	28.84	73.21	75.85	85.81
Refcoco	all	CIder	17,596	29.76	34.26	9.47	34.2	34.75	33.56
Refcoco+	all	CIder	7,578	28.92	31.01	9.05	31.82	32	30.66
Refcocoq	all	CIder	12,596	57.76	59.23	19.35	52.18	58.02	59.26
ScienceQA	scienceqa-img	Acc	2,017	70.43	72.88	28.56	70.15	73.57	81.85
SEED-Bench	Seed-1	Image-Acc	17,990	60.49	67.06	65.97	64.74	65.64	69.55
SEED-Bench-2	Seed-2	Acc	24,371	57.89	59.88	60.83	59.88	60.72	64.98
TextCaps	val	CIDEr	3,166	98.15	103.92	70.39	71.79	67.39	67.11
TextVQA	val	exact_match	5,000	46.07	48.73	65.76	64.85	66.92	69.31
VizWiz(val)	val	Acc	4,319	54.39	56.65	63.79	60.64	63.56	66.61
VQAv2	val	Acc	214,354	76.64	78.26	80.32	80.06	80.92	82.07

Table 28: List of websites selected for LIVEBENCH.

Name	URL	Category
BBC Main	<a href="https://www.bbc.com/">https://www.bbc.com/</a>	General News
BBC News	<a href="https://www.bbc.com/news">https://www.bbc.com/news</a>	News
BBC Sport	<a href="https://www.bbc.com/sport">https://www.bbc.com/sport</a>	Sports
BBC Business	<a href="https://www.bbc.com/business">https://www.bbc.com/business</a>	Business
BBC Innovation	<a href="https://www.bbc.com/innovation">https://www.bbc.com/innovation</a>	Innovation
BBC Culture	<a href="https://www.bbc.com/culture">https://www.bbc.com/culture</a>	Culture
BBC Travel	<a href="https://www.bbc.com/travel">https://www.bbc.com/travel</a>	Travel
BBC Future Planet	<a href="https://www.bbc.com/future-planet">https://www.bbc.com/future-planet</a>	Environment
CNN Main	<a href="https://edition.cnn.com/">https://edition.cnn.com/</a>	General News
CNN Politics	<a href="https://edition.cnn.com/politics">https://edition.cnn.com/politics</a>	Politics
CNN Entertainment	<a href="https://edition.cnn.com/entertainment">https://edition.cnn.com/entertainment</a>	Entertainment
CNN Style	<a href="https://edition.cnn.com/style">https://edition.cnn.com/style</a>	Style
Bloomberg Economics	<a href="https://www.bloomberg.com/economics">https://www.bloomberg.com/economics</a>	Economics
Bloomberg Industries	<a href="https://www.bloomberg.com/industries">https://www.bloomberg.com/industries</a>	Industries
Bloomberg Technology	<a href="https://www.bloomberg.com/technology">https://www.bloomberg.com/technology</a>	Technology
Bloomberg Politics	<a href="https://www.bloomberg.com/politics">https://www.bloomberg.com/politics</a>	Politics
Bloomberg Opinion	<a href="https://www.bloomberg.com/opinion">https://www.bloomberg.com/opinion</a>	Opinion
WSJ Main	<a href="https://www.wsj.com/">https://www.wsj.com/</a>	General News
WSJ Africa	<a href="https://www.wsj.com/world/africa?mod=nav_top_subsection">https://www.wsj.com/world/africa?mod=nav_top_subsection</a>	Africa
WSJ Americas	<a href="https://www.wsj.com/world/americas?mod=nav_top_subsection">https://www.wsj.com/world/americas?mod=nav_top_subsection</a>	Americas
WSJ Asia	<a href="https://www.wsj.com/world/asia?mod=nav_top_subsection">https://www.wsj.com/world/asia?mod=nav_top_subsection</a>	Asia
WSJ China	<a href="https://www.wsj.com/world/china?mod=nav_top_subsection">https://www.wsj.com/world/china?mod=nav_top_subsection</a>	China
WSJ Europe	<a href="https://www.wsj.com/world/europe?mod=nav_top_subsection">https://www.wsj.com/world/europe?mod=nav_top_subsection</a>	Europe
WSJ Middle East	<a href="https://www.wsj.com/world/middle-east?mod=nav_top_subsection">https://www.wsj.com/world/middle-east?mod=nav_top_subsection</a>	Middle East
WSJ India	<a href="https://www.wsj.com/world/india?mod=nav_top_subsection">https://www.wsj.com/world/india?mod=nav_top_subsection</a>	India
WSJ Oceania	<a href="https://www.wsj.com/world/oceania?mod=nav_top_subsection">https://www.wsj.com/world/oceania?mod=nav_top_subsection</a>	Oceania
WSJ Russia	<a href="https://www.wsj.com/world/russia?mod=nav_top_subsection">https://www.wsj.com/world/russia?mod=nav_top_subsection</a>	Russia
WSJ UK	<a href="https://www.wsj.com/world/uk?mod=nav_top_subsection">https://www.wsj.com/world/uk?mod=nav_top_subsection</a>	UK
WSJ Science	<a href="https://www.wsj.com/science?mod=nav_top_subsection">https://www.wsj.com/science?mod=nav_top_subsection</a>	Science
WSJ Archaeology	<a href="https://www.wsj.com/science/archaeology?mod=nav_top_subsection">https://www.wsj.com/science/archaeology?mod=nav_top_subsection</a>	Archaeology
WSJ Biology	<a href="https://www.wsj.com/science/biology?mod=nav_top_subsection">https://www.wsj.com/science/biology?mod=nav_top_subsection</a>	Biology
WSJ Environment	<a href="https://www.wsj.com/science/environment?mod=nav_top_subsection">https://www.wsj.com/science/environment?mod=nav_top_subsection</a>	Environment
WSJ Physics	<a href="https://www.wsj.com/science/physics?mod=nav_top_subsection">https://www.wsj.com/science/physics?mod=nav_top_subsection</a>	Physics
WSJ Space	<a href="https://www.wsj.com/science/space-astronomy?mod=nav_top_subsection">https://www.wsj.com/science/space-astronomy?mod=nav_top_subsection</a>	Space
WSJ Central Banking	<a href="https://www.wsj.com/economy/central-banking?mod=nav_top_subsection">https://www.wsj.com/economy/central-banking?mod=nav_top_subsection</a>	Central Banking
WSJ Consumers	<a href="https://www.wsj.com/economy/consumers?mod=nav_top_subsection">https://www.wsj.com/economy/consumers?mod=nav_top_subsection</a>	Consumers
WSJ Housing	<a href="https://www.wsj.com/economy/housing?mod=nav_top_subsection">https://www.wsj.com/economy/housing?mod=nav_top_subsection</a>	Housing
WSJ Jobs	<a href="https://www.wsj.com/economy/jobs?mod=nav_top_subsection">https://www.wsj.com/economy/jobs?mod=nav_top_subsection</a>	Jobs
WSJ Trade	<a href="https://www.wsj.com/economy/trade?mod=nav_top_subsection">https://www.wsj.com/economy/trade?mod=nav_top_subsection</a>	Trade
WSJ Global	<a href="https://www.wsj.com/economy/global">https://www.wsj.com/economy/global</a>	Global Economy
WSJ AI	<a href="https://www.wsj.com/tech/ai?mod=nav_top_subsection">https://www.wsj.com/tech/ai?mod=nav_top_subsection</a>	AI
WSJ Biotech	<a href="https://www.wsj.com/tech/biotech">https://www.wsj.com/tech/biotech</a>	Biotech
WSJ Cybersecurity	<a href="https://www.wsj.com/tech/cybersecurity?mod=nav_top_subsection">https://www.wsj.com/tech/cybersecurity?mod=nav_top_subsection</a>	Cybersecurity
WSJ Personal Tech	<a href="https://www.wsj.com/tech/personal-tech?mod=nav_top_subsection">https://www.wsj.com/tech/personal-tech?mod=nav_top_subsection</a>	Personal Tech
Reuters Main	<a href="https://www.reuters.com/">https://www.reuters.com/</a>	General News
Reuters Aerospace and Defense	<a href="https://www.reuters.com/business/aerospace-defense/">https://www.reuters.com/business/aerospace-defense/</a>	Aerospace and Defense
Reuters Autos and Transportation	<a href="https://www.reuters.com/business/autos-transportation/">https://www.reuters.com/business/autos-transportation/</a>	Autos and Transportation
Reuters Davos	<a href="https://www.reuters.com/business/davos/">https://www.reuters.com/business/davos/</a>	Davos
Reuters Energy	<a href="https://www.reuters.com/business/energy/">https://www.reuters.com/business/energy/</a>	Energy
Reuters Environment	<a href="https://www.reuters.com/business/environment/">https://www.reuters.com/business/environment/</a>	Environment
Reuters Finance	<a href="https://www.reuters.com/business/finance/">https://www.reuters.com/business/finance/</a>	Finance
Reuters Healthcare	<a href="https://www.reuters.com/business/healthcare-pharmaceuticals/">https://www.reuters.com/business/healthcare-pharmaceuticals/</a>	Healthcare
Reuters Media and Telecom	<a href="https://www.reuters.com/business/media-telecom/">https://www.reuters.com/business/media-telecom/</a>	Media and Telecom
Reuters Retail and Consumer	<a href="https://www.reuters.com/business/retail-consumer/">https://www.reuters.com/business/retail-consumer/</a>	Retail and Consumer
Reuters Future of Health	<a href="https://www.reuters.com/business/future-of-health/">https://www.reuters.com/business/future-of-health/</a>	Future of Health
Reuters Future of Money	<a href="https://www.reuters.com/business/future-of-money/">https://www.reuters.com/business/future-of-money/</a>	Future of Money
Reuters Take Five	<a href="https://www.reuters.com/business/take-five/">https://www.reuters.com/business/take-five/</a>	Analysis
Reuters World at Work	<a href="https://www.reuters.com/business/world-at-work/">https://www.reuters.com/business/world-at-work/</a>	World at Work
Reuters Breakingviews	<a href="https://www.reuters.com/breakingviews/">https://www.reuters.com/breakingviews/</a>	Opinion
Reuters Technology	<a href="https://www.reuters.com/technology/">https://www.reuters.com/technology/</a>	Technology
Reuters Cybersecurity	<a href="https://www.reuters.com/technology/cybersecurity/">https://www.reuters.com/technology/cybersecurity/</a>	Cybersecurity
Reuters Space	<a href="https://www.reuters.com/technology/space/">https://www.reuters.com/technology/space/</a>	Space
Reuters Disrupted	<a href="https://www.reuters.com/technology/disrupted/">https://www.reuters.com/technology/disrupted/</a>	Disruption
Reuters Momentum	<a href="https://www.reuters.com/technology/reuters-momentum/">https://www.reuters.com/technology/reuters-momentum/</a>	Technology
Reuters Investigations	<a href="https://www.reuters.com/investigations/">https://www.reuters.com/investigations/</a>	Investigations
Andreessen Horowitz	<a href="https://a16z.com/news-content/#latest">https://a16z.com/news-content/#latest</a>	Technology
Hacker News	<a href="https://news.ycombinator.com/">https://news.ycombinator.com/</a>	Technology
Reddit	<a href="https://www.reddit.com/?rdt=48006">https://www.reddit.com/?rdt=48006</a>	Social Media
Crunchbase News	<a href="https://news.crunchbase.com/">https://news.crunchbase.com/</a>	Startups
CCTV	<a href="https://www.cctv.com/">https://www.cctv.com/</a>	International News