Redundancy Principles for MLLMs Benchmarks

Zicheng Zhang[1,2]*, Xiangyu Zhao[1,2]*, Xinyu Fang[1,3], Chunyi Li[1,2], Xiaohong Liu[2], Xiongkuo Min[2], Haodong Duan[1]†, Kai Chen [1]†, Guangtao Zhai[1,2]†,
[1]Shanghai AI Laboratory, [2]Shanghai Jiaotong University, [3]Zhejiang University

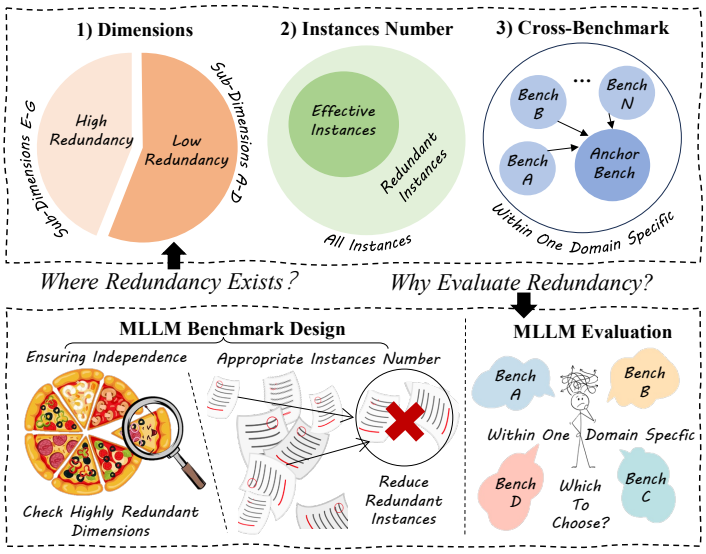


Figure 1: *Where Redundancy Exists?* and *Why Evaluate Redundancy?*

Abstract

With the rapid iteration of Multi-modality Large Language Models (MLLMs) and the evolving demands of the field, the number of benchmarks produced annually has surged into the hundreds. The rapid growth has inevitably led to significant redundancy among benchmarks. Therefore, it is crucial to take a step back and critically assess the current state of redundancy and propose targeted principles for constructing effective MLLM benchmarks. In this paper, we focus on redundancy from three key perspectives: **1)** Redundancy of benchmark capability dimensions, **2)** Redundancy in the number of test questions, and **3)** Cross-benchmark redundancy within specific domains. Through the comprehensive analysis over hundreds of MLLMs' performance across more than 20 benchmarks, we aim to quantitatively measure the level of redundancy lies in existing MLLM evaluations, provide valuable insights to guide the future development of MLLM benchmarks, and offer strategies to refine and address redundancy issues effectively.

1 Introduction

Model Evaluation has always played a crucial role in the development of Multi-modal Large Language Models (MLLMs). Benchmarks serve not only as tools for assessing model accuracy but also as catalysts for driving innovation and improvements within the field. In the early stages, traditional model evaluation benchmarks such as GQA (Hudson and Manning, 2019), VQA-V2 (Antol et al., 2015), VizWiz (Bigham et al., 2010), and TextVQA (Singh et al., 2019) are characterized by relatively simple questions and answers, with responses often being a single word. This limits the depth of understanding and reasoning required from the models, making them less effective at evaluating the complex capabilities of modern MLLMs that are expected to handle more nuanced tasks. With the emergence of more powerful MLLMs (Achiam et al., 2023; Team et al., 2023; Chen et al., 2024b; Wang et al., 2024b; Li et al., 2024a; Liu et al., 2024), traditional evaluation frameworks have become inadequate to meet the flexible evaluation requirements. In response, a new generation of VQA benchmarks has arisen, such as MMBench (Liu et al., 2025), MMVet (Yu et al., 2023a), and MMMU (Yue et al., 2024).

[1]Equal Contribution.
[2]Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12492–12504

As MLLMs have rapidly iterated and evolved, their diverse capabilities across various domains have garnered increasing attention, which has led to the development of specialized benchmarks to evaluate MLLMs' performance in specific areas, like Mathematics Task (Lu et al., 2023; Zhang et al., 2025; Wang et al., 2024a; Zou et al., 2024), Optical Character Recognition (OCR) (Mishra et al., 2019; Liu et al., 2023a; Mathew et al., 2021), Medical Field (Hu et al., 2024), Remote Sensing (Li et al., 2024e), Agents (Yang et al., 2024), GUIs (Baechler et al., 2024), and so on.

The rapid proliferation of benchmarks has inevitably introduced significant redundancies, with overlapping capabilities being assessed and recurring questions appearing within and across benchmarks. Such redundancies create inefficiencies in model evaluation, repeatedly testing similar aspects of MLLM performance without contributing meaningful new insights. Additionally, this trend risks overemphasizing certain task types while neglecting others, potentially distorting research priorities. In this work, we address these challenges through a comprehensive and systematic exploration.

1.1 Identifying Redundancy

Redundancy is an intrinsic and multifaceted issue in benchmarks, appearing in several key forms:

- **Redundancy across dimensions (intra-bench):** Tasks within the same benchmark may evaluate overlapping capabilities of MLLMs, leading to repetitive assessments.
- **Redundancy among instances (intra-bench):** Certain instances closely resemble others, providing minimal additional differentiation or insight for model evaluation.
- **Redundancy across benchmarks within specific domains:** Benchmarks targeting specific domains often exhibit overlapping objectives or scopes, resulting in duplicated efforts across different evaluation sets.

1.2 Ideal Redundancy Principles

Effective benchmarks should adhere to the following principles regarding redundancy:

- **Independence of dimensions:** Ideal benchmarks should ensure that its dimensions are largely independent, minimizing overlap between them. However, some degree of redundancy may be inevitable when certain capabilities naturally require the interaction of multiple foundational skills, and redundancy should be carefully balanced to avoid excessive overlap.
- **Optimal instance count:** A well-designed benchmark should strike a balance in the number of instances: neither too few nor too many, to ensure reliable and meaningful evaluations without introducing unnecessary redundancies.
- **Domain representativeness:** A comprehensive benchmark targeted to a specific domain should represent the domain. This may involve purposeful overlap with other benchmarks within the same domain to reflect shared core capabilities.

1.3 Benifits of Evaluating Redundancy

Evaluating and addressing redundancy offers several significant benefits, as shown in Fig. 1:

- **Optimizing benchmark design: 1).** Determines whether certain dimensions within a benchmark warrant separate assessments or can be consolidated; **2).** Identifies the minimal and sufficient number of instances required for accurate evaluation; **3).** Assesses the necessity of introducing new benchmarks within specific domains.
- **Enhancing efficiency in MLLM evaluation: 1).** Determines whether a benchmark deviates from the domain's distribution ; **2).** Identifies the anchor benchmarks required to evaluate model performance within the domain.

By systematically addressing redundancy, we can not only enhance the principles of benchmark design but also alleviate the resource demands of MLLM evaluation, creating a more streamlined and effective evaluation ecosystem.

2 Redundancy Framework

We present a framework for evaluating redundancy among MLLM capabilities, defined as specific tasks within a benchmark. Our framework is grounded in the following prior assumption:

When evaluating similar capabilities, the performance rankings of MLLMs should exhibit strong correlation. Conversely, significant differences in these rankings suggest the evaluated capabilities are relatively independent.

Based on this principle, we propose the **Performance Correlation Redundancy Framework**, which quantifies redundancy by measuring the correlation of MLLM performance rankings. To ensure robustness and generalization capability, we leverage the comprehensive data from VLMEvalKit (Duan et al., 2024), which includes

diverse benchmarks and performance results from more than 100 MLLMs.

2.1 Dimensions Redundancy

Assume a benchmark consists of a set of dimensions, denoted as $X = \{X_1, X_2, \ldots, X_m\}$, where each X_i represents a specific dimension. Let N denote the number of MLLMs evaluated on these dimensions. For a given dimension X_i, we denote the ranking of the N MLLMs on this dimension as R_i. To quantify the redundancy of X_i, we compute the average rank correlation between R_i and the rankings R_j of all other dimensions X_j ($j \neq i$). Formally, the redundancy $\rho(X_i)$ is defined as:

$$\rho(X_i) = \frac{1}{m-1} \sum_{\substack{j=1 \\ j \neq i}}^{m} \text{CORR}(R_i, R_j), \quad (1)$$

where $\text{CORR}(R_i, R_j)$ is the correlation coefficient between the rankings R_i and R_j.

- High $\text{CORR}(R_i, R_j)$ values can help identify potentially redundant dimension pairs.
- $\rho(X_i)$ represents the average redundancy level of dimension X_i, quantifying its overall overlap.

By calculating the redundancy $\rho(X_i)$ for all dimensions X_i in the benchmark and averaging these values, we can obtain the overall internal redundancy of the benchmark as well. Formally, the benchmark internal redundancy ρ_{BI} is defined as:

$$\rho_{BI} = \frac{1}{m} \sum_{i=1}^{m} \rho(X_i), \quad (2)$$

where $\rho(X_i)$ is the redundancy of the i-th dimension as previously defined. This metric reflects the average similarity among all dimensions within the benchmark. A lower ρ_{BI} suggests that the dimensions are relatively independent and diverse.

2.2 Instances Redundancy

Let a benchmark contain M total instances (*e.g.*, QA pairs). To evaluate redundancy, we begin by calculating the MLLM performance rankings obtained over the full set of all M instances, denoted as the ground-truth ranking R_{GT}. We then randomly sample a subset of the instances, comprising A% of the total M, and compute the corresponding MLLM rankings, denoted as R_{sample}. To quantify the redundancy of the benchmark at a sampling ratio of $A\%$, we calculate the correlation coefficient between R_{sample} and R_{GT}. This correlation reflects how representative the sampled subset is of the entire benchmark. To reduce the effect of randomness, the sampling process is repeated $T = 100$ times, and the average correlation result is recorded. We define the instance redundancy of the benchmark at sampling ratio $A\%$, denoted as $\rho(A\%)$, as follows:

$$\rho(A\%) = \frac{1}{T} \sum_{1 \leq t \leq T} \text{CORR}(R_{A\%_t}, R_{\text{GT}}), \quad (3)$$

where $R_{A\%_t}$ represents the MLLM ranking based on the sampled $A\%$ instances at the t_{th} time, and R_{GT} is the MLLM ranking based on the full M instances within the MLLM benchmark. The interpretation of $\rho(A\%)$ is straightforward:

- A higher $\rho(A\%)$ indicates that the sampled instances are highly representative of the entire benchmark, and the remaining $1 - A\%$ instances contribute little additional information.
- Conversely, a lower $\rho(A\%)$ suggests that the sampled instances are less representative, and more instances are needed to capture the variability of the full benchmark.

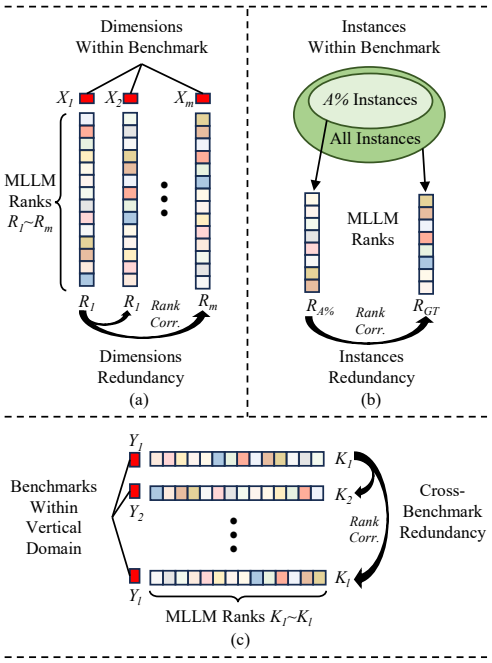
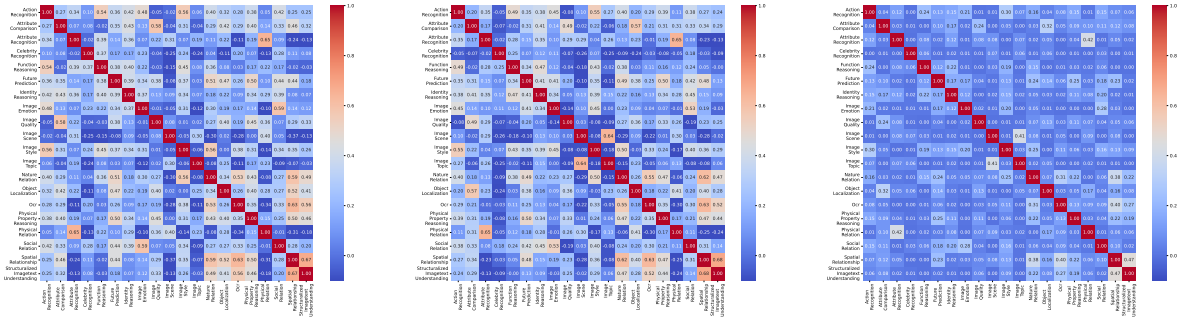


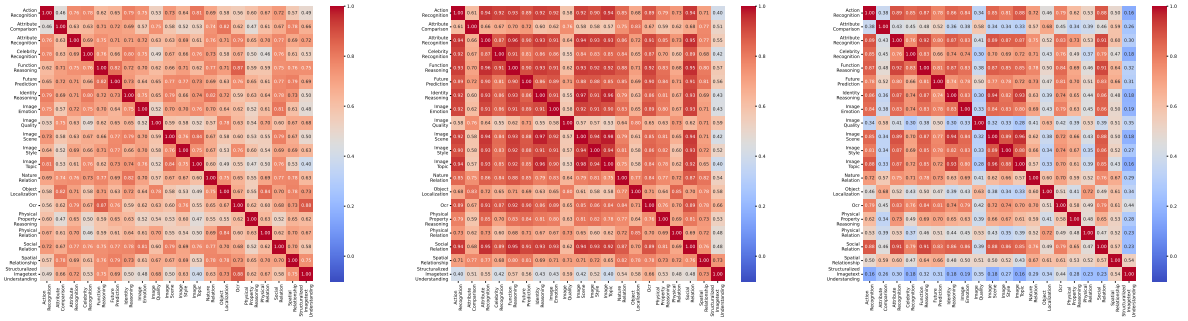
Figure 2: A quick look at the redundancy framework, where (a), (b), and (c) show the general process of computing dimensions redundancy, instances redundancy, and cross-benchmark redundancy respectively.

2.3 Cross-Benchmark Redundancy

Consider $Y = \{Y_1, Y_2, \ldots, Y_l\}$, a collection of l benchmarks within a specific domain (*e.g.*, object hallucination, visual reasoning, visual perception). Let N represent the number of MLLMs evaluated



(a) 50^{+} SRCC dimensions redundancy. (b) 50^{+} PLCC dimensions redundancy. (c) 50^{+} R2 dimensions redundancy.



(d) 50^{-} SRCC dimensions redundancy. (e) 50^{-} PLCC dimensions redundancy. (f) 50^{-} R2 dimensions redundancy.

Figure 3: Visualizations of dimensions redundancy for MMBench (Liu et al., 2025) on Top-50 and Bottom-50 (*marked as 50^{+} and 50^{-}*) MLLMs respectively. More benchmark results can be found in Appendix. B.

across these benchmarks. For a given benchmark Y_i, let K_i denote the ranking of the N MLLMs based on their performance on Y_i. To identify key anchor benchmarks within this domain (an anchor benchmark can serve as a representative over multiple other benchmarks), we focus on benchmarks that demonstrate high redundancy with others in the domain (Zohar et al., 2024). We define the redundancy of a benchmark $\rho(Y_i)$ as the average rank correlation coefficient between K_i and the rankings K_j of all other benchmarks Y_j ($j \neq i$) in the domain. Formally, $\rho(Y_i)$ is expressed as:

$$\rho(Y_i) = \frac{1}{l-1} \sum_{\substack{j=1 \\ j \neq i}}^{l} \text{CORR}(K_i, K_j), \qquad (4)$$

where $\text{CORR}(K_i, K_j)$ is the correlation coefficient between the rankings K_i and K_j. The interpretation of $\rho(Y_i)$ is as follows:

- A higher $\rho(Y_i)$ indicates that benchmark Y_i exhibits strong similarity with others in the domain, suggesting that it is highly representative of the domain's capabilities or evaluation focus.
- Conversely, a lower $\rho(Y_i)$ indicates that benchmark Y_i shares less overlap with others, implying that it is less redundant and may capture unique/distinct aspects of the domain, or incorporate noises that are not related to the domain.

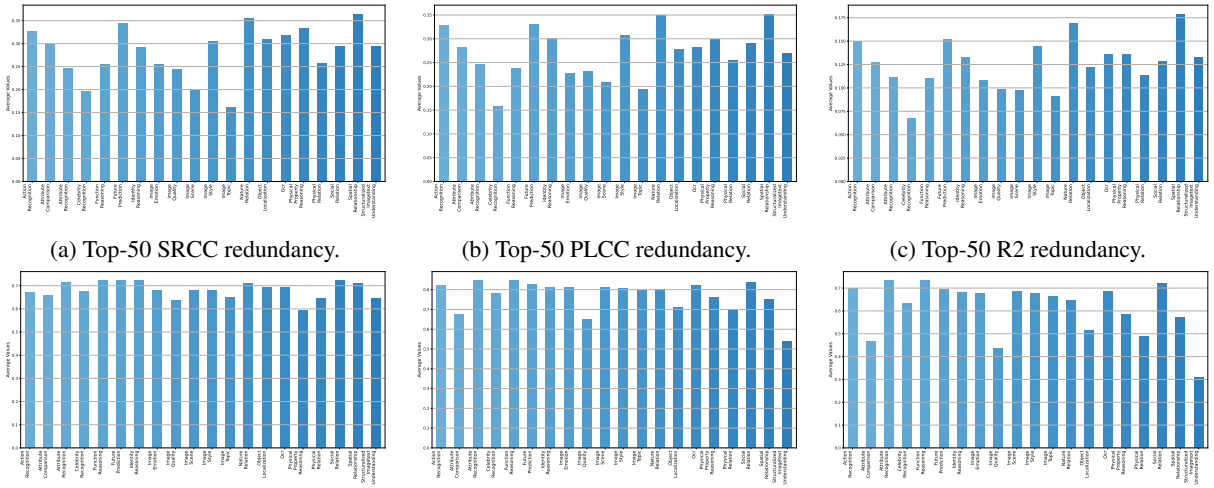
2.4 Correlation Metrics

In this work, we adopt multiple metrics to describe the correlation between two set of performance numbers, including the Spearman Rank Correlation Coefficient (SRCC), the Pearson Linear Correlation Coefficient (PLCC), and the R^2 Score (R-squared Coefficient of Determination).

- **SRCC** is an evaluation metric that measures rank similarity, capturing how well the relative order between two rankings aligns.
- **PLCC** quantifies linear similarity, assessing how closely the rankings follow a linear relationship.
- **R^2 Score**, on the other hand, evaluates the proportion of variance explained by the ranking relationship, serving as a measure of goodness-of-fit.

2.5 Top-K Analysis

Considering that the performance of top-tier MLLMs often garners greater attention on benchmarks, we can streamline the redundancy analysis by focusing only on the top-K MLLMs with the highest overall performance on a given benchmark, rather than incorporating all MLLMs in the calculation. By selecting the top-K models, we can better target the analysis of benchmark redundancy across different performance tiers. This approach also simplifies the process of maintaining and updating our framework as new MLLMs are introduced.



(a) Top-50 SRCC redundancy. (b) Top-50 PLCC redundancy. (c) Top-50 R2 redundancy.

(d) Bottom-50 SRCC redundancy. (e) Bottom-50 PLCC redundancy. (f) Bottom-50 R2 redundancy.

Figure 4: Bar plots of dimensions redundancy for MMBench (Liu et al., 2025) on Top-50 and Bottom-50 MLLMs. The redundancy values are computed by averaging the redundancy of each dimension with the redundancy of all other dimensions.

3 Experiment & Discussion

We use the evaluation results of hundreds of MLLMs obtained through the VLMEvalKit (Duan et al., 2024) as our data source for conducting experiments and analysis. All the data sources are open-sourced and available on HuggingFace [1].

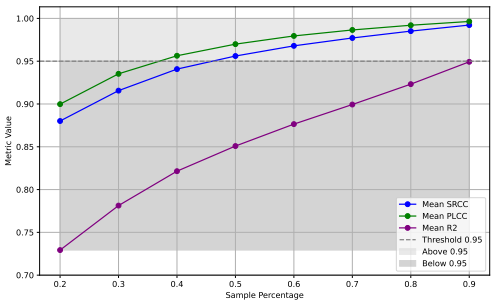
3.1 Exploring Dimension Redundancy

To comprehensively demonstrate the application of our redundancy framework in MLLM benchmarks, we conduct a detailed case study using the widely adopted and dimensionally diverse MMBench benchmark (v1.1) (Liu et al., 2025). We categorize the MLLMs into two groups, Top-50 and Bottom-50, based on their overall performance in MMBench. This categorization enables us to highlight the differences in redundancy exhibited by MMBench when evaluating MLLMs with varying levels of capability. The dimension redundancy results are illustrated in Fig. 3 and Fig. 4, from which we derived several interesting insights.

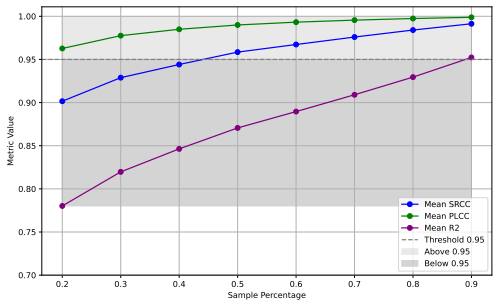
Top-50 Redundancy. Figs. 3a and 3b visually illustrate the redundancy of SRCC and PLCC across various sub-dimensions, allowing for a quick analysis of which dimensions exhibit high correlations. For example, the tasks **Image Emotion** and **Social Relation** display strong redundancy, suggesting a significant overlap in the skills they assess. Similarly, **Structuralized Image-Text Understanding** demonstrates notable redundancy with several other dimensions, such as **Spatial Relationship**, **Physical Property Reasoning**, **OCR**, and **Nature Relation**, indicating that these tasks collectively represent the diverse abilities required to perform **Structuralized Image-Text Understanding**. In contrast, **Image Topic** and **Image Scene** exhibit relatively low redundancy with other dimensions, as shown in Figs. 4a to 4c. This could arise from the inherent complexity of assessing the overall topic and scene of an image, which is often less correlated with evaluating specific attributes or relationships. For instance, strong performance in recognizing individual attributes does not necessarily imply a comprehensive understanding of the overall topic or scene. However, Fig. 3b reveals that these two dimensions exhibit redundancy in terms of PLCC, suggesting potential overlaps within certain contexts. Another interesting insight arises from **Celebrity Recognition**, a knowledge-based task that remains relatively independent of other dimensions, which primarily measure perceptual abilities. As a result, it consistently exhibits significantly lower redundancy across SRCC, PLCC, and R^2. Conversely, high levels of redundancy are observed for Nature Relation and Spatial Relationship, as shown in Figs. 4a to 4c. This is attributed to the fact that these two dimensions serve as fundamental skills required by numerous other tasks, making their overlap a cornerstone of the broader evaluation framework.

Bottom-50 Redundancy. The results for the Bottom-50 redundancy, as shown in Figs. 4d to 4f, reveal a striking trend where nearly all dimensions

[1] https://huggingface.co/datasets/VLMEval/OpenVLMRecords



(a) Instances redundancy with Top-50 MLLMs.



(b) Instances redundancy with Bottom-50 MLLMs.

Figure 5: Visualizations of average instance redundancy for (a) Top-50 MLLMs and (b) Bottom-50 MLLMs across 18 LMM benchmarks (A-Bench (Zhang et al., 2024a), AI2D (Kembhavi et al., 2016), BLINK (Fu et al., 2025), HallusionBench (Guan et al., 2023), MMBench (Liu et al., 2025), MMMU (Yue et al., 2024), MME (Fu et al., 2024), MMStar (Chen et al., 2024a), MMT (Ying et al., 2024), MMVet (Yu et al., 2023b), OCRBench (Liu et al., 2023b), Q-Bench (Wu et al., 2023; Zhang et al., 2024b), R-Bench-Dis (Li et al., 2024d), RealWorldQA (xAI, 2024), ScienceQA (Lu et al., 2022), SeedBench_IMG (Li et al., 2023), SeedBench2_Plus (Li et al., 2024b)). Notably, each data point represents the average of 100 sampling iterations to mitigate the impact of randomness.

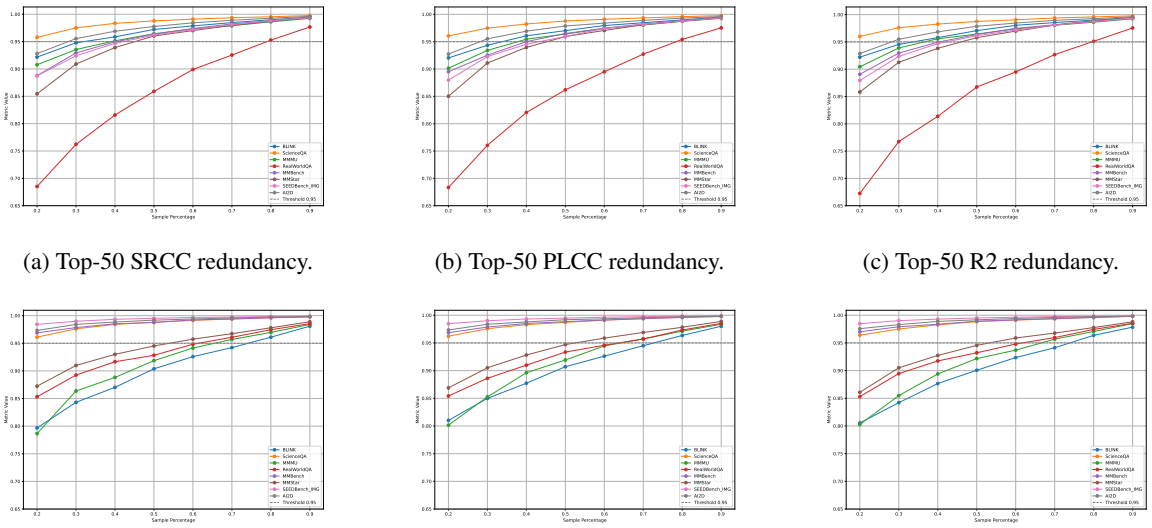
exhibit significantly higher redundancy compared to the Top-50 redundancy. Specifically, most dimension pairs achieve SRCC and PLCC scores exceeding 0.6 (Figs. 4d and 4e), leading to an interesting conclusion: **the dimensions appear to be more redundant for Bottom-50 MLLMs than for Top-50 MLLMs**. This phenomenon can primarily be attributed to the fact that Bottom-50 MLLMs generally underperform across all capabilities. For these models, as their foundational abilities improve, incremental enhancements in one dimension often drive simultaneous improvements across others. This results in high consistency in performance rankings across dimensions, thereby causing relatively high dimensional redundancy. In contrast, the Top-50 MLLMs have already achieved relatively strong foundational capabilities. Consequently, more complex tasks across different dimensions introduce greater variability, allowing for more differentiation between performance in those dimensions. This leads to noticeably lower levels of redundancy for the Top-50 models. These findings emphasize the importance of carefully selecting the MLLMs included in redundancy analysis. Specifically, avoiding models with universally poor performance is crucial to ensure that the evaluation yields meaningful and accurate insights.

3.2 Exploration Instance Redundancy

We include the evaluation results from 18 publicly available benchmarks in VLMEvalKit (Duan et al., 2024) in our experiments, with the average performance across benchmarks presented in Fig. 5. We adopt a similarity threshold of 0.95 for partitioning[2], This leads to an intriguing conclusion: **a majority of existing MLLM benchmarks exhibit significant redundancy in their instances when ranking both Top-50 and Bottom-50 MLLMs, with at least 50% of the instances being redundant.** This indicates that many benchmarks could reduce their instance counts by half without significantly affecting the ranking of MLLMs being tested. The R^2 score provides further insight, as it measures how effectively the final performance of MLLMs can be predicted using sampled instances. Compared to ensuring accurate ranking, achieving high accuracy in predicting the absolute performance of MLLMs requires a much larger number of instances. For example, both Top-50 and Bottom-50 MLLMs require over 90% of the instances to achieve an R^2 score greater than 0.95. This distinction highlights that fewer instances are sufficient for reliable ranking than for precise performance prediction.

We also compare redundancy tendencies between Top-50 and Bottom-50 MLLMs, as shown in Figs. 5a and 5b. Notably, at the same 0.95 threshold for SRCC and PLCC, Bottom-50 MLLMs require significantly fewer instances than Top-50 MLLMs. This implies that accurately ranking higher-performing MLLMs (Top-50) demands more instances, while ranking lower-performing MLLMs (Bottom-50) can be achieved with fewer

[2]Ranks with SRCC and PLCC coefficients exceeding 0.95 are considered nearly identical, with only marginal differences in very few cases (Hauke and Kossowski, 2011).



(a) Top-50 SRCC redundancy. (b) Top-50 PLCC redundancy. (c) Top-50 R2 redundancy.

(d) Bottom-50 SRCC redundancy. (e) Bottom-50 PLCC redundancy. (f) Bottom-50 R2 redundancy.

Figure 6: Benchmark-specific instance redundancy for (a) Top-50 MLLMs and (b) Bottom-50 MLLMs. The benchmarks include BLINK (Fu et al., 2025), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), RealWorldQA (xAI, 2024), MMBench (Liu et al., 2025), MMStar (Chen et al., 2024a), SeedBench_IMG (Li et al., 2023), and AI2D (Kembhavi et al., 2016). The selection of the Top-50 and Bottom-50 MLLMs is based on the corresponding benchmark.

instances. Consequently, the redundancy of benchmark instances correlates strongly with the capability of the MLLMs being evaluated: **the stronger the MLLMs, the lower the redundancy of the benchmark instances**.

From the benchmark-specific results (Fig. 6), the redundancy gap between Top-50 and Bottom-50 MLLMs remains consistent across different benchmarks. Further examination reveals considerable variation in redundancy levels between benchmarks. For example, in the Top-50 redundancy analysis, RealWorldQA (xAI, 2024) demonstrates relatively low redundancy, requiring nearly 80% of the instances to reach saturation, while other benchmarks require far fewer. However, for Bottom-50 MLLMs, redundancy levels across benchmarks increase significantly, and the differences between them narrow. This illustrates that benchmark redundancy is more prominent when evaluating less capable MLLMs.

It is important to note that the conclusions above are based on the statistical analysis of mainstream benchmarks. Specialized benchmarks, with unique design goals or tasks, require case-by-case analyses to assess their instance redundancy accurately. Therefore, while these results provide general insights into redundancy trends for standard benchmarks, further evaluation is necessary for niche or task-specific benchmarks.

3.3 Exploring Cross-Benchmark Redundancy

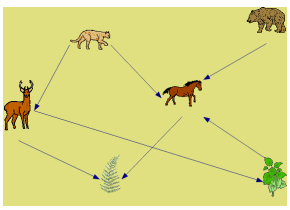
To analyze cross-benchmark redundancy, we focus on the Math domain, specifically examining several popular mathematical benchmarks: MathVista (Lu et al., 2023), MathVision (Zhang et al., 2025), MathVerse (Wang et al., 2024a), and DynaMath (Zou et al., 2024). We utilize the available evaluation results of 37 MLLMs listed on the OpenCompass Reasoning Leaderboard[3] and assess their ranking performance across these math benchmarks. The corresponding heatmap is presented in Fig. 8. The results reveal that, although all four benchmarks are designed to evaluate the mathematical abilities of MLLMs, the correlations between them are not particularly strong. Among them, MathVista (Lu et al., 2023) exhibits the least redundancy, showing the lowest correlation with the other benchmarks. In contrast, MathVerse and MathVision demonstrate high redundancy, indicating a strong correlation with other benchmarks. These differences suggest varying levels of overlap in their evaluation focus areas.

To better understand the variability across benchmarks, we analyze their task distributions. While MathVerse and MathVision are focused on standard mathematical tasks, resulting in the highest correla-

[3] https://huggingface.co/spaces/opencompass/Open_LMM_Reasoning_Leaderboard

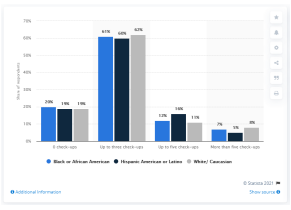
Q: Was this a square pizza?
A: no

(a)



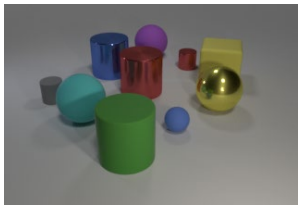
Q: In the diagram of the food web shown what will most directly be affected by the loss of the trees?
A: horses

(b)



Q: What is the lowest value in blue bar?
A: 7

(c)



Q: Subtract all tiny purple shiny cubes. Subtract all large purple balls. How many objects are left?
A: 9

(d)

Figure 7: Examples of tasks excluded from the MathVista benchmark. (a), (b), and (c) showcase tasks derived from the *general-vqa* category, including *Scientific Figure Understanding*, *General VQA*, and *Chart/Table/Diagram QA*. Panel (d) presents questions extracted from the CLEVR dataset but categorized as *math-targeted-vqa*.

tion and substantial overlap with other benchmarks, MathVista includes 30%-40% of questions outside traditional mathematics, such as tasks related to *Scientific Figure Understanding*, *General VQA*, and *Chart/Table/Diagram QA* (see Fig. 7(a)(b)(c) for examples). As discussed in Sec. 2.3, low redundancy can arise from unique elements specific to a domain or from irrelevant tasks, which we consider "noise" within the dataset. For instance, general VQA tasks, while broadly useful, have limited relevance to assessing mathematical ability and contribute to this noise. To quantify the impact, we remove general VQA tasks from MathVista and recalculate its redundancy with other benchmarks. After this refinement, the redundancy between MathVista and other mathematical benchmarks significantly increases, aligning more closely with their task profiles. Additionally, we identify and exclude *CLEVR*-derived questions categorized as *math-targeted vqa* within MathVista, which also have limited relevance to mathematical capabilities (examples in Fig. 7(d)). This further increases overlap with specialized mathematical benchmarks, demonstrating that removing irrelevant tasks improves alignment and reduces noise.

Therefore, we propose the following principles for benchmark design within a domain:

- A benchmark intended to broadly assess model performance in one domain should demonstrate relatively high redundancy with other in-domain benchmarks, reflecting comprehensive coverage of diverse sub-capabilities.
- A specialized benchmark should display lower redundancy with other benchmarks, focusing on distinct capabilities to fill the vacancy, complement broader assessments, and provide a unique perspective on specific topics in a domain.

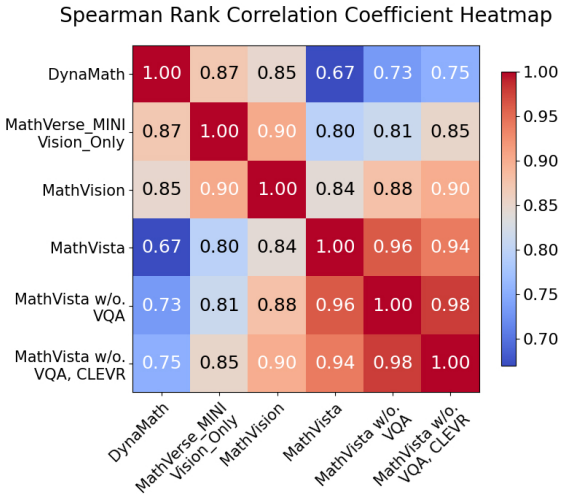


Figure 8: Cross-benchmark redundancy map. MathVision and MathVersion are more focused on the core domain of mathematics (with relatively higher redundancy across other math benchmarks), making them more suitable for benchmarking the mathematical capabilities of MLLMs in a narrow sense.

4 Conclusion

In conclusion, this paper addresses the pervasive issue of redundancy in MLLM benchmarks, impacting both the effectiveness and efficiency of model evaluation. We identify redundancy at three levels: dimension, instance, and cross-benchmark redundancy, and propose a framework with actionable guidelines to improve benchmark design. By promoting the independence of dimensions, optimizing instance counts, and ensuring purposeful redundancy within specific domains, our framework streamlines evaluations and enhances reliability. Case studies further demonstrate its utility in refining current practices, paving the way for more efficient and accurate MLLM assessments.

5 Limitations

The limitations of this work is as follows:

- The assumption that MLLM performance rankings should show strong correlation when evaluating similar capabilities may not always hold. In some cases, performance on seemingly similar tasks could diverge due to subtle task differences, domain-specific nuances, or differences in model strengths.
- The use of correlation metrics (SRCC, PLCC, and R^2) to quantify redundancy may be limited in capturing the full complexity of model performance across different tasks and domains. These metrics may not adequately account for differences in task difficulty, model behavior under various conditions, or the impact of outliers.
- The redundancy value is not fixed when using different selections of MLLMs for calculation. This bias could result in misleading conclusions about the redundancy or uniqueness of certain benchmarks.

6 Acknowledgements

This research was partly supported by grants of Science and Technology Commission of Shanghai Municipality (STCSM, Grant No. 20DZ1200203, 2021SHZDZX0102), National Key R&D Program of China (No.2022ZD0161600), the Shanghai Postdoctoral Excellence Program (No.2023023), China Postdoctoral Science Fund (No.2024M751559), National Natural Science Foundation of China (NSFC, Grant No. 62301310, 623B2073), and Shanghai Artificial intelligence Laboratory.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

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.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

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.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024c. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. 2024d. R-bench: Are your large multimodal model robust to real-world corruptions? *arXiv preprint arXiv:2410.05474*.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. 2024e. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *arXiv preprint arXiv:2406.12384*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023a. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023b. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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. Realworldqa dataset. Available at https://huggingface.co/datasets/xai-org/RealworldQA.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023a. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2025. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. 2024a. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 2024b. Q-bench: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. 2024. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*.

A Metrics Equation

To evaluate the consistency and accuracy of predictions, we employ three widely used metrics: the Spearman Rank Correlation Coefficient (SRCC), the Pearson Linear Correlation Coefficient (PLCC), and the Coefficient of Determination (R^2). These metrics provide complementary perspectives on model performance, capturing rank-based, linear, and variance-explained relationships, respectively. The mathematical definitions are detailed below.

1) The SRCC measures the rank-based relationship between predicted and true values. It is defined as:

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where:

$$d_i = \text{rank}(x_i) - \text{rank}(y_i),$$

and n is the number of data points. A higher SRCC indicates a stronger monotonic relationship between the rankings of predicted and ground truth values.

2) The PLCC quantifies the linear relationship between predicted and true values. It is computed as:

$$\text{PLCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where:

- x_i and y_i are the data points,
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

A higher PLCC indicates a stronger linear relationship between predicted and ground truth values.

3) The R^2 score represents the proportion of variance in the ground truth values that is explained by the predictions. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where:

- y_i are the ground truth values,
- $\hat{y}_i$ are the predicted values,
- $\bar{y}$ is the mean of the ground truth values.

An R^2 score closer to 1 indicates a better fit between the predictions and the ground truth.

B Extra Dimensions Redundancy Maps

We present the dimension redundancy maps for AI2D (Kembhavi et al., 2016) and SEED-Bench (Li et al., 2024c), as shown in Fig. 9 and Fig. 10.

1. **Key Observations from the Redundancy Maps:**
   - In Fig. 9, it is evident that the dimension 'lifeCycles' exhibits the highest redundancy, particularly with 'typesOf'.
   - Similarly, in Fig. 10, the 'Instance Identity' dimension shows the highest redundancy and is most closely related to 'Scene Understanding'.
2. **Trends in Top-50 vs. Bottom-50 Redundancy:**
   - A clear pattern emerges when comparing the Top-50 and Bottom-50 redundancy maps. Nearly all Bottom-50 dimensions display significantly higher redundancy than their Top-50 counterparts. This observation supports our conclusion that **dimensions tend to exhibit greater redundancy for Bottom-50 MLLMs compared to Top-50 MLLMs**.
   - This phenomenon can be attributed to the overall underperformance of Bottom-50 MLLMs across various capabilities. As these models begin to improve, enhancements in their foundational abilities often lead to simultaneous progress across multiple dimensions. This results in a high degree of similarity in performance rankings, contributing to elevated dimensional redundancy.
   - In contrast, Top-50 MLLMs already possess relatively strong foundational capabilities. As a result, more challenging tasks across different dimensions introduce greater differentiation, reducing redundancy and creating more distinct performance profiles.
3. **Implications for Redundancy Analysis:**
   - To ensure a reasonable and accurate evaluation during redundancy analysis, it is crucial to exclude MLLMs with consistently poor performance. Including such models could skew the analysis by disproportionately inflating redundancy, as their universal underperformance does not provide meaningful insights into inter-dimensional relationships.

C Redundancy Practice Recommendations

To ensure benchmarks are reliable and efficient, we recommend incorporating redundancy detection into the benchmark design process after its initial testing on a set of MLLMs. This critical step identifies potential redundancies across dimensions/instances/cross-benchmark overlaps, leading to more precise and meaningful evaluations.

1. **Dimension Redundancy Check.**

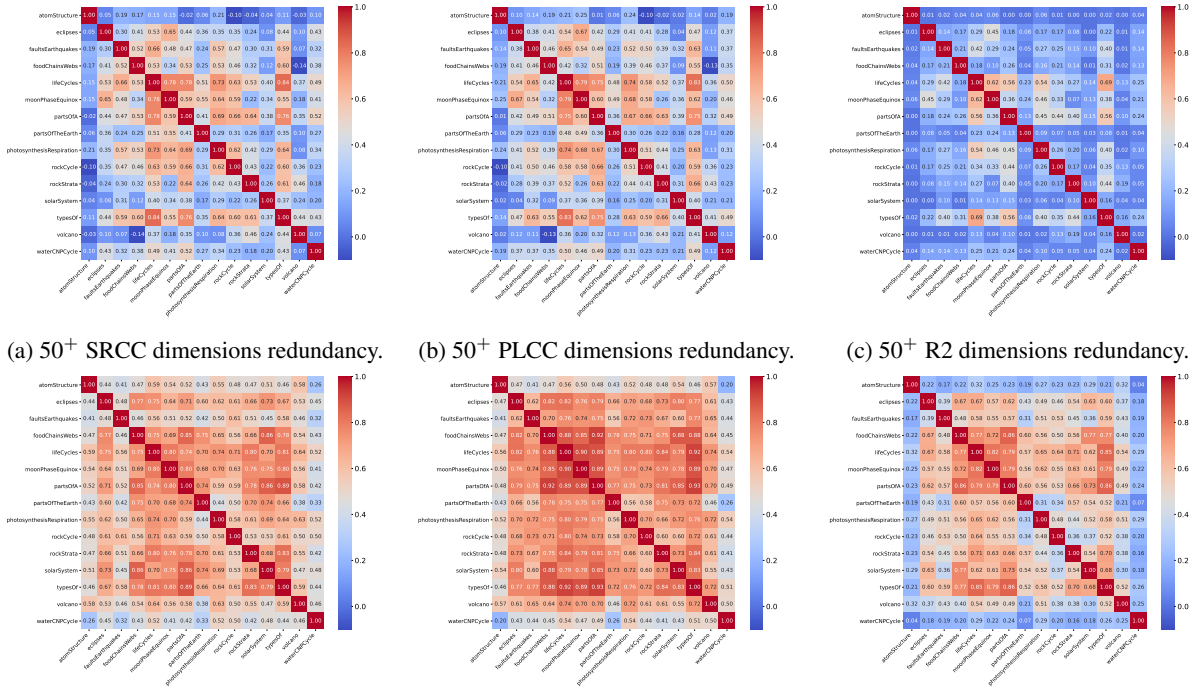
   Calculate the dimensional redundancy within the benchmark, with particular attention to dimensions exhibiting overall high redundancy. Analyze the redundancy heatmap to identify pairs of dimensions with exceptionally strong correlations, as these may indicate overlapping capabilities being assessed. For such cases, evaluate whether these dimensions are truly necessary or whether they assess similar or redundant skills.

2. **Instance Redundancy Check.**

   Compute the instance redundancy curve to determine whether a smaller subset of benchmark instances can produce results comparable to the full instance set. If significant instance redundancy is identified, the benchmark should be reviewed, and redundant instances should be reduced. This not only streamlines the evaluation process but also optimizes resource usage without compromising the accuracy of results.

3. **Cross-benchmark Redundancy Check.**

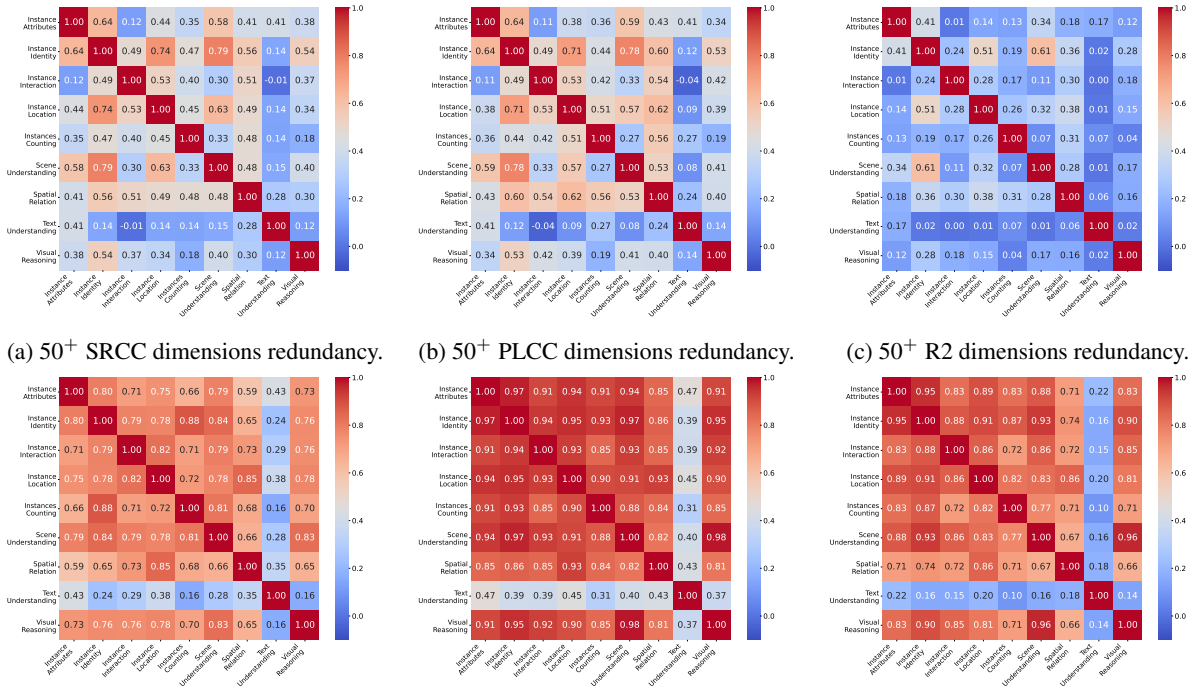
   If the benchmark is intended to serve as a representative for a specific domain, measure its cross-benchmark redundancy relative to other benchmarks within the domain. Higher redundancy indicates stronger representativeness, making it a reliable choice for tasks requiring domain coverage. Conversely, if the goal is to fill a vacancy in the specific domain (*e.g.*, focusing on a specific topic in mathematics that is not covered by previous benchmarks) maintaining low redundancy is a more favorable choice. For use cases focusing on core capabilities within a specific domain under limited resources, it is recommended to select the benchmark with the highest cross-benchmark redundancy. This ensures that the benchmark comprehensively covers the essential skills while minimizing unnecessary overlaps.



(a) 50^+ SRCC dimensions redundancy. (b) 50^+ PLCC dimensions redundancy. (c) 50^+ R2 dimensions redundancy.

(d) 50^- SRCC dimensions redundancy. (e) 50^- PLCC dimensions redundancy. (f) 50^- R2 dimensions redundancy.

Figure 9: Visualizations of dimensions redundancy for AI2D (Kembhavi et al., 2016) on Top-50 and Bottom-50 MLLMs (*marked as 50^+ and 50^-*).



(a) 50^+ SRCC dimensions redundancy. (b) 50^+ PLCC dimensions redundancy. (c) 50^+ R2 dimensions redundancy.

(d) 50^- SRCC dimensions redundancy. (e) 50^- PLCC dimensions redundancy. (f) 50^- R2 dimensions redundancy.

Figure 10: Visualizations of dimensions redundancy for SEED-Bench (Li et al., 2024c) on Top-50 and Bottom-50 MLLMs (*marked as 50^+ and 50^-*).