SAME IMAGE, DIFFERENT RANKING: TASK FORMULATION MATTERS IN MLLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are typically evaluated using benchmark scores that are often interpreted as stable measures of model capability. In this work we demonstrate inconsistencies in the VLM/MLLM performance measures in popular benchmarks. Instead, rankings tend to align with benchmark identity. That is, tasks within a single benchmark, even those targeting different capabilities, yield similar model rankings. Crucially, we identify *task formulation*, the textual definition of the task input and output format, as a source of such discrepancies. To study this, we construct controlled variants of benchmarks, holding the visual content and semantic goal fixed for each item, while varying the task formulation (binary, multiple-choice, or open-ended). We demonstrate that this change can substantially alter both model success on the same example and the resulting overall model rankings. A natural response is to evaluate tasks under multiple formulations, but this substantially increases evaluation costs. To mitigate this, we propose a *formulation-aware subset selection* based on Item Response Theory (IRT). This approach efficiently approximates multi-formulation evaluation using compact subsets, which we confirm recover the performance trends as the full matrix. We conclude that reliable MLLM evaluation must account for variation across task formulations, and offer an effective and efficient protocol to do so.



Input Image

Task Formulation: Binary
Q: Is there only one real apple in this picture?
Model Prediction: Yes ✓
Task Formulation: Multiple-Choice (MCQ)
Q: How many real apples are in this picture?
(A) Four (B) Two **(C) One** (D) Info missing (E) None
Model Prediction: (C) ✓
Task Formulation: Open-Ended
Q: How many real apples are in the picture?
Model Prediction: "There are **two** real apples in the picture." ✗

Figure 1: **Evaluation Instability in MLLMs.** The same semantic task can yield different outcomes across formulations. The image and underlying semantic target are fixed while the textual task formulation varies. Models may answer the resulting formulations differently, illustrating that evaluation outcomes depend not only on the visual content and capability being tested, but also on how the task is formulated.

1 INTRODUCTION

Benchmark scores are the default mechanism for evaluating progress in multimodal large language models (MLLMs). Suites such as MME, MMBench, and MM-Vet (Fu et al. (2025); Liu et al. (2024a); Yu et al. (2024)) are routinely used to compare models, implicitly assuming that tasks intended to measure a specific *capability* (e.g., OCR, spatial reasoning) serve as stable, interchangeable units of evaluation. Under this assumption, leaderboard rankings are interpreted as reliable evidence of relative model quality for that underlying capability (Team (2024); HuggingFace (2024)).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we demonstrate that this assumption is frequently violated in multimodal evaluation, revealing inconsistencies in model rankings. To make this precise, we distinguish between a *benchmark*, which defines an evaluation suite; a *task*, which specifies an individual evaluation problem; its intended *capability*, which describes the ability it is designed to measure; and its *task formulation*, namely the textual input and output structure used to pose and evaluate it.

Through an extensive analysis of a model-by-task performance matrix spanning more than 250 MLLMs and multiple multimodal benchmarks, we find that model rankings are more strongly associated with benchmark identity than with the capability being evaluated (Section 2.2). Factor Analysis (FA) Fabrigar & Wegener (2011) confirms this structural divergence: tasks from the same benchmark tend to share model-ordering structure, whereas tasks targeting the same capability across different benchmarks do to a much lesser degree. Thus, capability labels alone do not define a transferable unit for comparing models: the model ordering induced by one benchmark need not transfer to another benchmark targeting the same capability. This benchmark-driven structure stands in stark contrast to recent findings in uni-modal LLM evaluation, where tasks targeting the same capability tend to share a latent organization across datasets Maimon et al. (2026) and LLM benchmark performance can be explained by a small set of latent capabilities. This divergence raises a natural question: what aspects of MLLM benchmark design contribute to this instability?

We identify *task formulation* as one primary source of the observed instability. While the visual input and semantic goal of a task may remain identical, different formulations, such as binary, multiple-choice, and open-ended questions, require substantially different prediction behaviors. To isolate this effect, we constructed controlled formulation variants of examples from MME, MMBench, and OCRBench Fu et al. (2025); Liu et al. (2024a;b), allowing only the textual formulation to vary while keeping the image and semantic goal fixed (Section 3).

Across all three benchmarks, changing the task formulation alone substantially alters model rankings. Agreement between rankings induced by different formulations, measured using Kendall's τ, is often low and can approach zero. The effect is also visible at the instance level: the same model can succeed under one formulation and fail under another for the same image and semantic goal. Together, these results demonstrate that model rankings are conditional on task formulation rather than invariant measures of capability.

Since benchmark leaderboards are routinely used to compare MLLMs, instability in model rankings undermines their role as reliable evidence for relative model performance on a desired capability.A natural response is to evaluate models across multiple formulations rather than relying on a single fixed formulation. However, evaluating every example under multiple formulations substantially increases computational and API costs. We therefore propose a *formulation-aware IRT-based subset-selection* method (Section 4) Lord (1980) that identifies informative examples while accounting for formulation-dependent variability. We show that compact subsets, comprising approximately 3–5$\%$ of the available examples, can recover the ranking structure of substantially larger multi-formulation evaluations, while explicit formulation coverage improves robustness across evaluation conditions.

Our core contributions are threefold: (1) We demonstrate, across multiple multimodal suites, that model rankings align more strongly with benchmark identity than with the underlying capabilities the benchmarks intend to measure. (2) We isolate task formulation as a critical source of this instability, showing that varying the task formulation while holding visual and semantic content fixed substantially alters relative model rankings. (3) We introduce a formulation-aware IRT subset selection method, demonstrating that compact subsets (3–5% of examples) can accurately recover the full ranking structure of massive multi-formulation evaluations.

We release the formulation-augmented versions of MME, MMBench, and OCRBench, together with our evaluation and subset-selection code.

2 Evaluation Setup and Benchmark-Dependent Model Rankings

A central challenge in multimodal evaluation is determining whether individual benchmark tasks provide a consistent basis for comparing models. To investigate this, we reconstruct a unified model-by-task performance matrix from standardized predictions across multiple benchmarks.

2.1 EVALUATION SETUP

Model-by-task performance matrix. For each model $m \in \mathcal{M}$ and task $t \in \mathcal{T}$, we compute a normalized performance score $p_{m,t} \in [0, 1]$ using the task's standard evaluation procedure, forming the model-by-task performance matrix: $\Pi \in [0, 1]^{|\mathcal{M}|\times|\mathcal{T}|}$, where each entry $\Pi_{m,t}$ denotes the performance of model m on task t. This representation allows us to compare the model-ordering structure induced by different tasks and, in particular, to test whether tasks targeting the same capability produce similar rankings across benchmarks.

Prediction data and scoring. We reconstruct task-level performance directly from raw prediction files using the VLMEVALKIT framework Team (2024), ensuring consistent parsing and scoring procedures across the evaluation suite. For categorical, binary, and multiple-choice tasks, predictions are evaluated against the task's reference answers using the corresponding benchmark-defined metric, such as accuracy or F1. Scores are normalized to $[0, 1]$ and averaged across examples. For free-form generative tasks, correctness is scored on a 0-9 scale using DEEPSEEK-REASONER DeepSeek-AI (2025) as a judge, then normalized to $[0, 1]$ and averaged across examples. While LLM-based evaluation can introduce noise Zheng et al. (2023), our findings remain consistent when restricting analysis to classification tasks.

Benchmark suite. We evaluate over 250 MLLMs across a diverse suite of benchmarks, including MME Fu et al. (2025), MMBench Liu et al. (2024a), MM-Vet Yu et al. (2024), OCRBench Liu et al. (2024b), HallusionBench Liu et al. (2023), AI2D Kembhavi et al. (2016), and MathVista Lu et al. (2024). Because these benchmarks contain tasks targeting overlapping capabilities (e.g., OCR, spatial reasoning), we can directly compare model rankings across datasets. For the controlled formulation experiments in Section 3, we utilize a fixed subset of 16 modern MLLMs.

Why benchmark identity matters. The benchmarks we use differ not only in the capabilities they target, but also in image distributions, annotation procedures, task formulations, and evaluation protocols. Consequently, two tasks intended to measure a similar capability may differ substantially in how questions are phrased, how responses are structured, and how answers are evaluated.

2.2 BENCHMARK-DEPENDENT MODEL RANKINGS

Recent work demonstrates that LLM benchmark performance can be explained by a small number of latent capability factors, where tasks targeting the same capability induce similar model rankings regardless of the dataset Maimon et al. (2026). We ask whether this capability-centered organization holds for MLLMs, or if variations in image distributions, task formulations, and evaluation protocols cause model-ordering to align more strongly with benchmark identity. We see that **Benchmark Identity Explains Substantial Variation Beyond Capability Labels**.

Factor Analysis. To answer this question, we analyze the global structure of model performance using Principal Axis Factoring (PAF) Fabrigar & Wegener (2011); Bartholomew & Knott (1999), following the same methodology used in Maimon et al. (2026). PAF identifies latent factors that explain shared variance among tasks, allowing us to test whether tasks induce similar model-ordering behavior (see Appendix A for details).

If model rankings primarily reflect shared capabilities, tasks intended to evaluate the same capability should load on similar latent factors regardless of the benchmark from which they originate. Instead, Figure 2 (representative subset) shows a different pattern. Beyond a dominant general factor reflecting overall model strength (Appendix B.4), most benchmarks are associated with a distinct latent factor[1]. Tasks from the same benchmark generally load strongly onto the same specific factor, while tasks targeting similar capabilities across different benchmarks do not. Thus, similarities in model-ordering behavior are more strongly associated with benchmark membership than with the intended capability.

Consistency Ranking. To quantify this observation, we compared ranking consistency within each benchmark against a null distribution of randomly sampled tasks.

[1]Figure 2 shows loadings after rotation where the shared variance reflected in the first unrotated eigenvalue is distributed across the rotated factors.

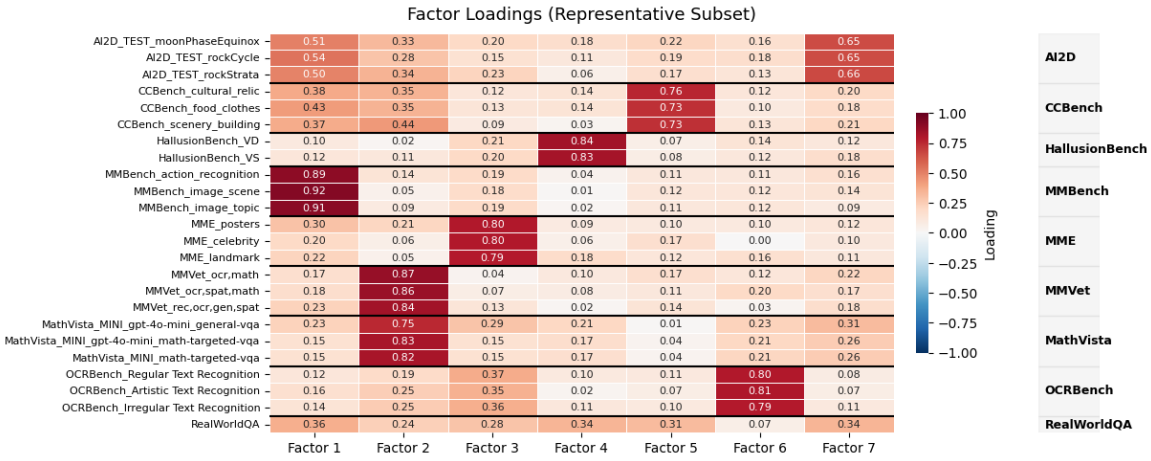


Figure 2: PAF loadings for a representative subset of tasks. Rows represent individual tasks—grouped by their source benchmark on the far right—and columns represent the extracted latent factors. Most benchmarks are associated with a dominant latent factor, indicating that model-ordering behavior is primarily organized by benchmark rather than by shared capability labels. Full matrix in Appendix B.1

Table 1: Cross-benchmark agreement for representative task pairs. Low Kendall's τ indicates that capability labels (e.g., OCR) do not guarantee consistent model rankings across benchmarks

Task Pair	Pearson	Spearman	Kendall τ
MME OCR vs. *OCRBench* Text Recognition	0.34	0.18	0.12
OCRBench Text Recognition vs. *RealWorldQA*	0.38	0.48	0.33
MME Celebrity vs. *MMBench* Celebrity	0.50	0.49	0.34
MME Position vs. *MMBench* Localization	0.67	0.66	0.50

Tasks from the same benchmark exhibit substantially greater ranking consistency than expected under the null distribution (Fig. 9), confirming strong benchmark-dependent structure in model rankings.

These results establish strong benchmark-dependent structure, but do not identify which benchmark properties are responsible for it. Benchmarks differ in image distributions, annotation procedures, response formats, and evaluation protocols, any of which could contribute to the observed structure. We therefore next test whether tasks intended to measure the same capability nevertheless produce different model rankings across benchmarks (further elaboration in B.6).

2.3 Cross-Benchmark Disagreement

To isolate this benchmark-dependent variation, we compare task pairs across different benchmarks intended to evaluate the identical capability (e.g., OCR, localization) but differing in formulation or evaluation procedure. We measure ranking agreement using Pearson, Spearman, and Kendall's τ Kendall (1938) correlations over the model-performance vectors.

As shown in Table 1, agreement is severely limited. Even closely related tasks yield low Kendall's τ (e.g., 0.12), and no task pair reaches consistent alignment. These results show that **task labels/names/definitions alone don't define a transferable unit of evaluation**: tasks labeled as targeting similar capabilities can induce substantially different model rankings across benchmarks (Further benchmark-level analysis in Appendix C.2, D). This severe lack of agreement challenges the implicit transferability assumption of modern leaderboards: that proficiency in a capability on one benchmark guarantees it elsewhere. Capability labels alone do not define a transferable evaluation unit.

Summary. Both latent-factor and direct ranking analyses show model orderings are far more consistent within individual benchmarks than across tasks targeting the same capability. Therefore,

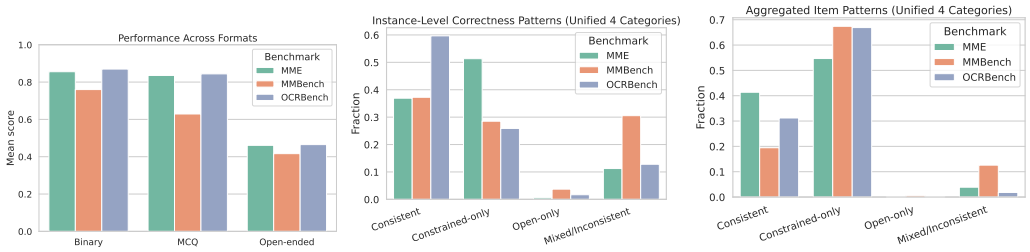


Figure 3: **(a)** Mean performance across formulations. **(b)** Instance-level correctness patterns. **(c)** Item-level correctness patterns.

capability labels do not guarantee transferable model rankings across benchmarks. Next (3), we isolate task formulation as one source of this instability.

3 Task Formulation as a Source of Evaluation Instability

Section 2.2 showed that model rankings often vary across benchmarks, even when tasks target similar capabilities. We now ask whether *task formulation* (textual input & output structure through which a task is presented) can itself alter model rankings, while holding the visual input and semantic goal fixed. To test this, we construct controlled formulation variants and measure their effects on aggregate performance, model rankings, and instance-level correctness.

3.1 Setup: Controlled Formulation Variation

We study formulation effects using three widely used VLM benchmarks: MME (Fu et al., 2025), MMBench (Liu et al., 2024a), and OCRBench (Liu et al., 2024b).

For each benchmark, we construct Binary, Multiple-Choice (MCQ), and Open-ended formulations while preserving the original image and semantic goal. This enables controlled comparison across formulations while minimizing changes to task content. We use 3 distinct LLMs for augmentation and randomly sample among their outputs to minimize generator specific phrasing artifacts. To rule out evaluation artifacts, all formats are scored by the LLM judge and cross-checked against exact match. On 500 human-annotated items, the judge agrees with human labels in 92% of cases (Cohen's κ=0.83). Full details are in Appendix E. We evaluate 16 modern MLLM spanning both proprietary and open-source systems.

3.2 Aggregate Effects Across Benchmarks

We first ask whether changing the task formulation while holding the visual input and semantic target fixed is sufficient to alter model rankings.

Figure 3(a) reports average performance across formulations and benchmarks. Across all three benchmarks, constrained formulations (Binary and MCQ) consistently yield higher success rates than Open-ended evaluation despite preserving the underlying visual content and semantic goal.

Table 2 shows that rank agreement remains low across all benchmarks, well below within-format agreement between disjoint item samples (last row). Several formulation pairs exhibit near-zero or even negative correlations, indicating that models that perform well under one formulation may rank substantially lower under another (more in Appendix F.2).

These correlations are comparable to, and often lower than, the cross-benchmark correlations in Section 2.2: changing the task formulation, with images and semantic goals fixed, alters rankings as much as switching to an entirely different benchmark. The effect is particularly striking because images and semantic goals remain fixed. Ranking instability therefore cannot be attributed solely to differences in visual content or capability coverage.

Table 2: Kendall τ across formulations (95% bootstrap CIs over items). *Within-format*: agreement between disjoint item samples of the same formulation. *: significantly below within-format (Appendix F.2).

Formulation Pair	MME	MMBench	OCRBench
Binary vs. MCQ	$0.27 \pm 0.15^*$	$0.19 \pm 0.11^*$	$-0.05 \pm 0.18^*$
Binary vs. Open	$0.10 \pm 0.22^*$	$-0.16 \pm 0.19^*$	$0.01 \pm 0.19^*$
MCQ vs. Open	$0.25 \pm 0.16^*$	$0.25 \pm 0.17^*$	0.36 ± 0.14
Within-format	0.78 ± 0.09	0.71 ± 0.10	0.58 ± 0.15

Structural Effects of Formulation Beyond aggregate rankings, we investigate whether formulation influences the structure of model behavior.

We report formulation-conditioned correlation analyses for all three benchmarks. Across benchmarks, tasks sharing the same formulation exhibit stronger behavioral similarity than tasks sharing semantic content but differing in formulation (Appendix I).

These results provide structural evidence that formulation influences the correlation structure of model rankings, complementing the benchmark-dependent effects observed in Section 2.2.

3.3 Instance-Level Mechanism

To understand how ranking instability emerges, we analyze correctness at the level of individual examples. For model m and item i, we record correctness under each formulation: $/c^{\text{bin}}_{m,i}, \quad c^{\text{mcq}}_{m,i}, \quad c^{\text{open}}_{m,i} \in \{0, 1\}$.

Figure 3(b) reports instance-level correctness patterns, while Figure 3(c) aggregates these patterns at the item level. Across all benchmarks, correctness is frequently preserved within constrained formulations but changes substantially when moving to Open-ended evaluation. The dominant formulation-sensitive pattern corresponds to success under Binary and MCQ but failure under Open-ended evaluation. This observation should not be interpreted as evidence that constrained formulations provide a more faithful estimate of model capability or that open-ended evaluation is inherently preferable or inferior. Rather, it demonstrates that changing the task formulation systematically changes which examples are solved correctly, and consequently alters the relative ordering of models. Detailed analyses in Appendix G show that this asymmetry persists across models, task categories, and benchmark families.

Constrained formulations reduce the effective answer space, whereas Open-ended formulations require free-form generation. These differences systematically change which examples individual models solve correctly, leading to different aggregate model rankings. Importantly, our results do not imply that any single formulation provides the 'correct' ranking. Instead, they show that model comparisons depend on the chosen evaluation formulation.

Interpretation. Across MME, MMBench, and OCRBench, changing task formulation consistently alters model predictions and the resulting model rankings. The effect appears at multiple levels: aggregate performance, model rankings, task-level structure, and instance-level correctness.

Together, these results indicate that task formulation is a systematic source of evaluation variability rather than a benchmark-specific artifact. Evaluation under a single fixed formulation therefore provides only a partial characterization of model performance, motivating evaluation procedures that account for formulation variability while remaining computationally efficient.

4 Efficient Evaluation via Formulation-Aware Subset Selection

Sections 2.2 and 3 show that model rankings depend on task formulation. A direct implication is that reliable VLM evaluation should compare models across multiple formulations rather than relying on a single fixed setup. However, evaluating every item under multiple formulations substantially increases evaluation cost.

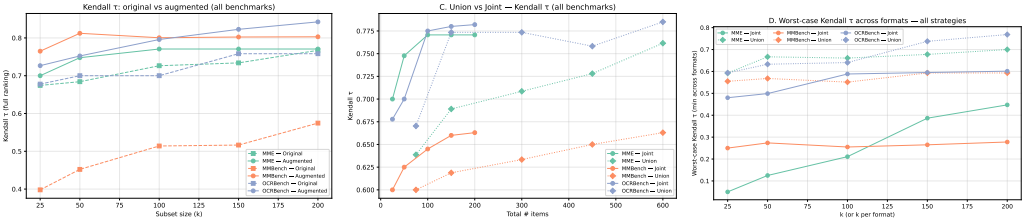


Figure 4: Subset selection performance across MME, MMBench, and OCRBench. **Left:** Ranking fidelity vs. subset size for original and augmented benchmarks. **Middle:** Overall ranking fidelity for joint and union subsets. **Right:** Worst-case formulation ranking fidelity for joint vs. union subsets.

We proceed in four steps: (1) we introduce an IRT-based *formulation-aware subset selection* framework (4.1); (2) we measure ranking and score fidelity across subset sizes (4.2); (3) we evaluate the efficiency–robustness trade-off among joint, formulation-specific, and union subsets (4.3); and (4) we evaluate the benefit of formulation-aware and cross-formulation generalization (4.4, 4.5).

Our goal is therefore not to replace multi-formulation evaluation, but to make it computationally feasible: to identify compact subsets that preserve the ranking information obtained from substantially larger formulation-augmented evaluations.

4.1 Subset Selection Framework

We aim to select a compact subset of items that preserves relative model rankings. Since discriminative items better distinguish models, we use Item Response Theory (IRT) (Lord, 1980; Baker, 2001), where each item is characterized by difficulty and discrimination parameters (2PL). To account for formulation effects, we allow item difficulty to vary across formulations. For formulation $f \in \mathcal{F}$, the probability that model m correctly answers item i is $P\big(c_{m,i}^{(f)} = 1\big) = \sigma\big(a_i(\theta_m - b_{i,f})\big)$, where θ_m is a latent model parameter used only to score item informativeness[2], a_i is item discrimination, and $b_{i,f}$ is the difficulty of item i under formulation f. Absorbing formulation variance into $b_{i,f}$ relaxes standard IRT's unidimensionality assumption. Subsets of size k maximize mean Fisher information across formulations, $\frac{1}{|\mathcal{F}|}\sum_f I_{i,f}(\hat{\theta})$) (details in Appendix J).

We evaluate subsets using three complementary criteria: (1) **Ranking fidelity:** agreement between model rankings on the full dataset and on the subset, measured by Kendall's τ; (2) **Score fidelity:** error in absolute performance estimates, measured by MAE; and (3) **Robustness:** worst-case ranking agreement across formulations, $\tau_{\min} = \min_{f\in\mathcal{F}} \tau_f$. The framework operates directly on the model-by-item performance matrix and is benchmark- and model-agnostic.

4.2 Subset Size and Compression

Figure 4 (left) shows that ranking fidelity improves rapidly with subset size on all three benchmarks, for both original and formulation-augmented data. With just 5% of the items, joint IRT selection reaches $\tau \approx 0.8$ with the full ranking, consistently outperforming random selection (Appendix J). Subsets also transfer to unseen models: in 5-fold cross-validation holding out three models from IRT estimation, held-out models deviate by 1.4 rank positions on average from their full-data rank (vs. 5.3 expected under random placement). This result shows that multi-formulation evaluation doesn't require evaluating all available examples: despite the added variability, the global ranking structure remains highly compressible.

In contrast, score fidelity improves more gradually: MAE decreases steadily but more slowly than ranking error (Appendix K), indicating that absolute scores are more sample-intensive to estimate than relative ordering. The gap is largest on formulation-augmented benchmarks, where scores aggregate over formulations and are therefore more sensitive to subset composition.

[2]Since θ_m is shared across formulations, the probability captures formulation-dependent difficulty shifts but not rank reversals. Rankings are therefore computed from observed responses on the subset, and robustness to reversals is measured via $\tau_{\min}$ and addressed by union subsets

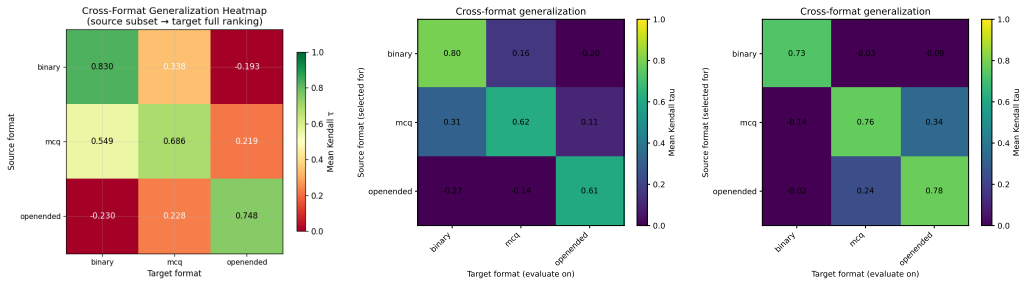


Figure 5: Cross-formulation generalization. Subsets perform best on the formulation used for selection and degrade when evaluated on other formulations.

4.3 Efficiency vs. Robustness Trade-off

We compare three subset construction strategies: (i) **Joint subsets:** a single subset selected across all formulations (joint selection doesn't assume a shared model ability; it seeks shared ranking information across formulations). (ii) **Formulation-specific subsets:** separate subsets optimized for each formulation. (iii) **Union subsets:** the union of formulation-specific subsets.

Figure 4 (middle) compares joint and union subsets. Joint subsets achieve strong overall ranking fidelity using substantially fewer samples. This indicates that average ranking structure can be captured efficiently by a compact shared subset.

However, in Figure 4 (right), under worst-case evaluation, joint subsets show lower fidelity across formulations, whereas union subsets maintain higher minimum agreement across all three benchmarks. This highlights a trade-off: joint subsets are more compact and preserve average rankings by integrating signals across formulations, demonstrating that formulation diversity drives construct validity, while union subsets require more samples but better preserve formulation-specific performance.

4.4 Benefit of Formulation-Aware Selection

We next compare formulation-aware IRT to standard IRT, which ignores formulation-dependent item behavior. Appendix J shows that formulation-aware IRT consistently improves worst-case ranking fidelity across MME, MMBench, and OCRBench (e.g., $\tau_{\min}$ on MMBench rises from -0.37 to 0.26 at $k{=}100$; Fig. 27). The improvement is particularly important on benchmarks with strong formulation sensitivity, where standard IRT may select items that preserve average rankings but fail under specific formulations. By allowing item difficulty to vary across formulations, formulation-aware IRT selects items that better preserve evaluation robustness. These results show that modeling formulation-dependent variability is useful not only for diagnosing evaluation instability, but also for constructing more reliable evaluation subsets.

4.5 Cross-Formulation Generalization

Figure 5 shows that subsets selected for one formulation generalize imperfectly to others. Performance is highest on the formulation used for selection and degrades when evaluated on others. This confirms that different formulations induce partially distinct evaluation signals. It also explains why compact joint subsets can recover average rankings while still underperforming in worst-case evaluation: they do not explicitly guarantee coverage of formulation-specific structure. Additional analyses of subset behavior across task categories are provided in Appendix L.

Summary. If model rankings depend on task formulation, robust comparison requires evaluating across multiple formulations. Formulation-aware subset selection efficiently approximates this process. Across MME, MMBench, and OCRBench, compact subsets recover much of the full ranking structure, while additional formulation-specific coverage improves worst-case robustness. The resulting trade-off makes multi-formulation evaluation more practical: small joint subsets provide efficient ranking estimates, while larger union subsets ensure robustness across formulations

5 Related Work

Multimodal benchmark evaluation. Benchmarks such as MME, MMBench, and MM-Vet Fu et al. (2025); Liu et al. (2024a); Yu et al. (2024) are widely used to evaluate MLLMs, with aggregate scores and rankings often interpreted as measures of model ability. Prior work has highlighted dataset bias, distribution shift, contamination, and limited benchmark transferability Torralba & Efros (2011); Recht et al. (2019). We instead study whether model rankings remain stable across tasks targeting similar capabilities.

Prompt sensitivity and evaluation design. Prior work has shown that LMs can be sensitive to prompt phrasing and interaction format Webson & Pavlick (2022), primarily at the level of individual predictions and textual variation. In contrast, we study how task formulation in MLLM evaluation affects *relative model rankings*, showing that changing formulation alone can substantially alter model comparisons even when visual content and semantic goals are preserved.

Psychometrics and item selection. We use FA Fabrigar & Wegener (2011) and IRT Lord (1980) to analyze evaluation structure and select compact subsets that preserve model rankings. While IRT has been applied to LLM benchmark estimation using anchor sets and prompt templates (tinyBenchmarks; Maia Polo et al., 2024a; PromptEval; Maia Polo et al., 2024b), we apply it to multimodal evaluation, optimizing for ranking preservation and worst-case robustness across formulations.

6 Conclusion

This paper investigates the reliability of multimodal evaluation through standard benchmarks, demonstrating that model performance on a given capability is not invariant; it is highly conditioned on the textual *task formulation* and various benchmark-specific design characteristics. Consequently, aggregate benchmark scores should not be interpreted as a single, formulation-independent measure of model capability. In this work we show that leaderboard rankings are inherently conditional; changing the textual task formulation while keeping the visual input and semantic target intact can substantially change absolute success rates and change relative model rankings.

This sensitivity arises because different formulations of the same query probe fundamentally distinct aspects of model behavior. While constrained output formats (such as binary and multiple-choice) systematically simplify the prediction by restricting the output range, open-ended formulations demand generative synthesis and free-form instruction compliance. As a result, a single evaluation formulation provides only a partial characterization of model behavior.

To remedy this, evaluating a task across a spectrum of formulations offers a more robust assessment alternative, but doing so scales poorly due to increased computational and API costs. We mitigate this by proposing a *formulation-aware subset selection framework* based on Item Response Theory (IRT). We explicitly model format-dependent item difficulty, allowing one to reconstruct overall model rankings with high fidelity using only a small fraction (3–5%) of the total benchmark items.

Our overall empirical findings thus advocate for a paradigm shift in how multimodal leaderboards are constructed and interpreted. Rather than treating benchmark-induced model rankings as an absolute measure of capability, the community should treat evaluations as a multi-conditional profiling process. Rather then aiming for a single static score, aim for systematic characterization of model behavior across diverse task formulations and evaluation conditions.

7 Limitations

Our study has several limitations. First, we consider only three task formulations (binary, multiple-choice, and open-ended), and open-ended evaluation relies on an LLM-based judge Zheng et al. (2023), although our findings persist under exact-match scoring of constrained formats (with human labels validation). Second, while controlled experiments involve 16 MLLMs, bootstrapping and the 250-model analysis support robustness, though larger model pools would yield tighter estimates. Finally, we isolate task formulation as one source of evaluation instability. Future work should examine its interaction with dataset-specific biases and other evaluation factors.

AI USE STATEMENT

In this work, we used generative AI tools for synthetic data generation and as an automated evaluation judge. Specifically, multiple LLMs were employed to generate plausible distractors and formulate prompt variants for benchmark augmentation, as described in Section 3 and Appendix E. Additionally, an LLM judge was utilized to score open-ended responses, validated against human annotations. We also used generative AI tools to assist in editing and polishing the manuscript text. We have not used generative AI for drafting paper sections, proving mathematical claims, or research ideation. All AI-generated synthetic items and model evaluations were manually verified on representative subsets by the authors, and we take full responsibility for the final content, claims, and artifacts of this work.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Our study relies exclusively on publicly available multimodal benchmark datasets and models. The synthetic task augmentations introduced in this paper are derived strictly from these existing public assets and do not introduce private personal data or harmful content. Human validation was conducted solely on public benchmark examples to assess annotation and judge consistency, presenting no ethical concerns.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, all formal methodologies, including the PAF procedure (Appendix A) and the formulation-aware 2PL IRT subset selection framework (Appendix J), are mathematically detailed in the paper. The data generation protocols, prompt templates, and LLM judge specifications used for benchmark augmentation are described in Section 3 and Appendix E. Anonymized source code implementing the evaluation pipelines, subset selection algorithms, and the formulation-augmented benchmark datasets are provided in the supplementary material and will be made publicly available upon release.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Jinze Bai et al. Qwen-vl: A versatile vision-language model. *arXiv preprint arXiv:2308.12966*, 2023.

Frank B. Baker. *The Basics of Item Response Theory*. ERIC Clearinghouse, 2001.

David J. Bartholomew and Martin Knott. *Latent Variable Models and Factor Analysis*. Oxford University Press, 1999.

Raymond B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1 (2):245–276, 1966.

Lin Chen et al. Internvl: Scaling up vision foundation models and aligning for multimodal tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Google DeepMind. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Leandre R. Fabrigar and Duane T. Wegener. *Exploratory Factor Analysis*. Oxford University Press, 2011.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2): 179–185, 1965.

HuggingFace. Hugging face lmsys leaderboard. https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard, 2024.

Aniruddha Kembhavi et al. Ai2d: Interpreting scientific diagrams. *arXiv preprint arXiv:1603.07396*, 2016.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2(3):9, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Routledge, 1980. doi: https://doi.org/10.4324/9780203056615.

Urbano Lorenzo-Seva and Jos M. F. ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2):57–64, 2006.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *International Conference on Machine Learning (ICML)*, 2024a.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Aviya Maimon, Amir DN Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. From benchmarks to skills: Low-rank factors for llm evaluation, 2026. URL https://arxiv.org/abs/2507.20208.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

OpenCompass Team. Vlmevalkit: An open-source toolkit for evaluating vision-language models. https://github.com/open-compass/VLMEvalKit, 2024.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pp. 1521–1528, USA, 2011. IEEE Computer Society. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995347. URL https://doi.org/10.1109/CVPR.2011.5995347.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts?, 2022. URL https://arxiv.org/abs/2109.01247.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, pp. 57730–57754. PMLR, 2024.

Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

A Principal Axis Factoring Details

Principal Axis Factoring (PAF) is an exploratory factor analysis method Fabrigar & Wegener (2011); Bartholomew & Knott (1999) that decomposes shared variance among observed variables. Given a correlation matrix R, PAF seeks a decomposition

$$R = \Lambda\Lambda^\top + \Psi,$$

where Λ represents factor loadings and Ψ is a diagonal matrix of uniqueness terms.

Unlike PCA, which explains total variance, PAF focuses on common variance shared across variables. The estimation proceeds by iteratively refining communalities and extracting latent factors.

In this work, we apply PAF to the task correlation matrix to identify latent dimensions of shared model-ranking behavior. These latent factors are not tied to explicit capability labels, supporting the analysis in Section 2.2.

Rotation. We apply rotation to the extracted factors, following Maimon et al. (2026). The rotation allows factors to correlate, which is appropriate here because benchmark-specific factors are expected to be correlated through overall model strength. As a result, the general factor evident in the unrotated eigenvalues is not isolated as a single rotated factor but distributed across factors. All reported loadings are rotated loadings.

B Robustness of Benchmark-Conditioned Structure

B.1 Full PAF Loading Matrix

Figure 6 shows the full PAF loading matrix across all tasks.

Interpretation. The full loading matrix confirms that the benchmark-conditioned structure observed in the main text is not an artifact of selecting a representative task subset. Across the full task set, tasks from the same benchmark consistently co-load on shared latent factors.

B.2 Normalization Robustness

We evaluate the robustness of the findings in Section 2.2 to different normalization strategies. In addition to the default per-task z-score normalization used in the main experiments, we consider:

- per-task min–max normalization,
- per-benchmark z-score normalization.

For each normalization scheme, we rerun the full analysis pipeline, including factor analysis, benchmark-conditioned consistency, and benchmark-level ranking correlations.

B.3 Summary of Results

Table 3 summarizes the agreement between each normalization scheme and the baseline.

Overall, factor structure, benchmark consistency ordering, and cross-benchmark ranking relationships remain stable across normalization choices.

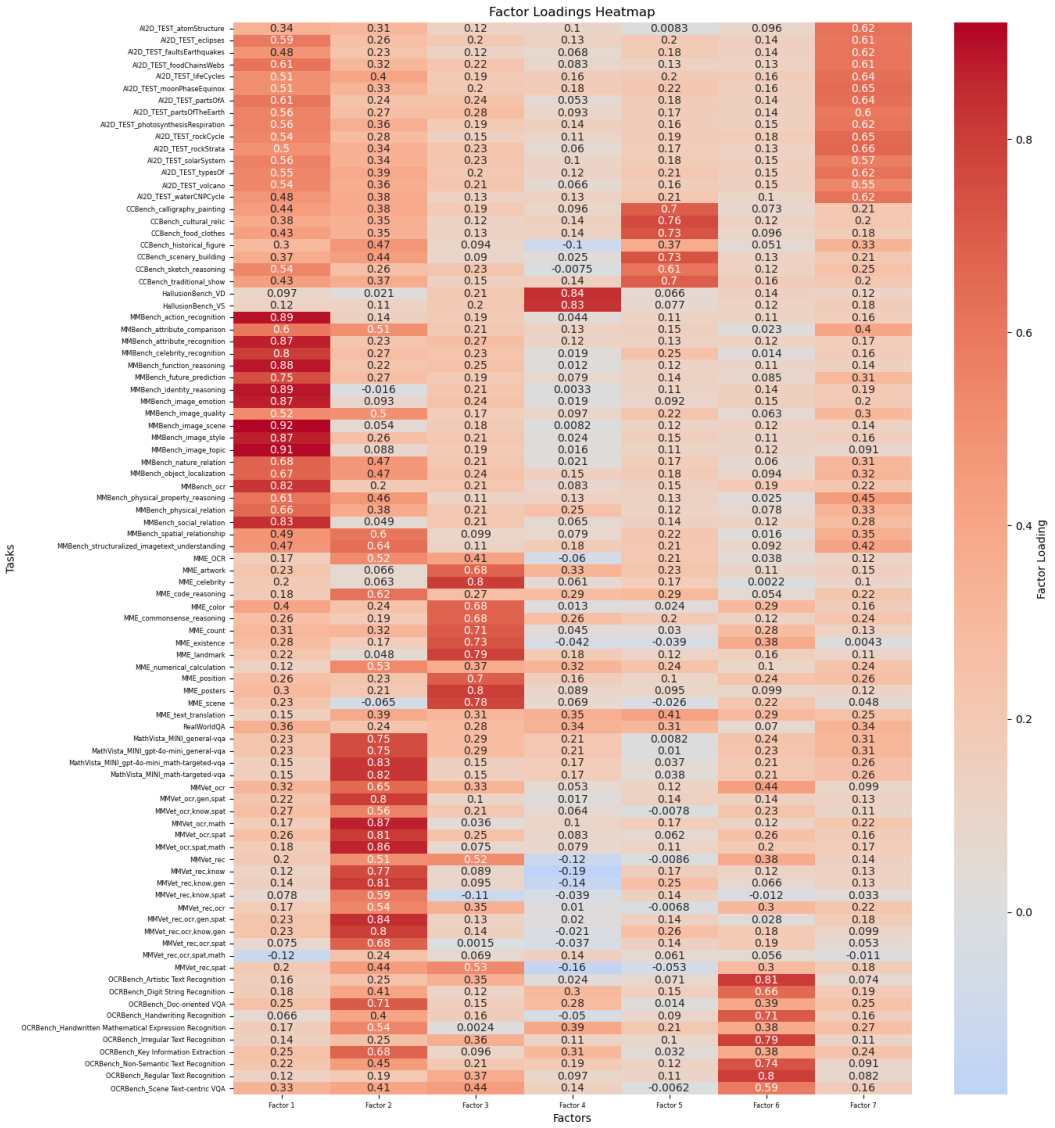


Figure 6: Full PAF loading matrix across all tasks. The same benchmark-conditioned structure observed in the main text is visible at full resolution: tasks from the same benchmark consistently co-load on shared latent factors. Compared to the representative subset shown in Figure 2, this visualization provides complete coverage of all tasks.

Table 3: Agreement between normalization schemes and the baseline. Factor structure, benchmark consistency ordering, and cross-benchmark correlations remain highly stable.

Normalization	Mean Factor Congruence	Kendall (Consistency Order)	Kendall (Benchmark Matrix)
Task z-score (baseline)	1.00	1.00	1.00
Task min–max	≈ 0.99	≈ 1.00	≈ 0.98
Benchmark z-score	≈ 0.97	≈ 1.00	≈ 0.95

B.4 Factor Stability

We evaluate robustness of the factor structure across factor counts and model subsets.

Factor count selection. To determine the appropriate number of factors, we use both scree analysis Cattell (1966) and parallel analysis Horn (1965). Figure 7 shows the scree plot of eigenvalues, while Figure 8 compares observed eigenvalues to a random baseline. The first eigenvalue ($\approx$49 across $\sim$90 tasks) indicates a dominant general factor: stronger models tend to perform well across all tasks. Our claim concerns the structure beyond this factor. The additional factors retained by parallel analysis align with benchmark identity rather than with capability labels (Figure 2).

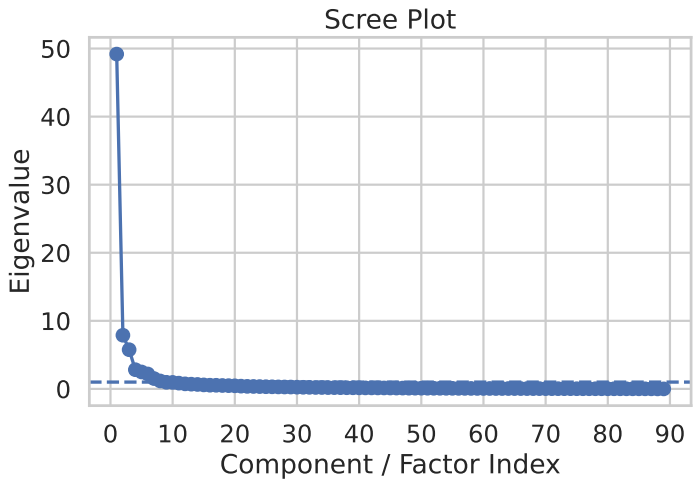


Figure 7: Scree plot of eigenvalues.

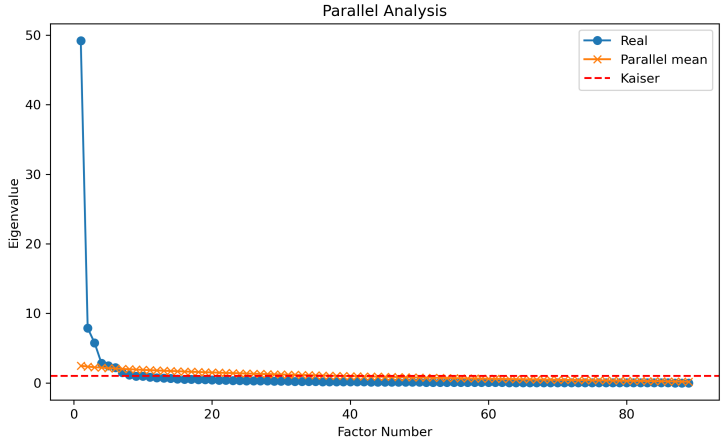


Figure 8: Parallel analysis comparing observed eigenvalues to random baseline.

Stability across factor counts. We evaluate the robustness of the factor structure by fitting PAF models with varying numbers of factors $k \in \{5, 6, 7, 8, 9\}$ and computing congruence between neighboring solutions using Tucker's congruence coefficient Lorenzo-Seva & ten Berge (2006). Table 4 reports the resulting congruence scores.

Across all tested values of k, mean congruence remains above 0.96, indicating that the dominant latent structure is stable across a reasonable range of factor counts.

Model subsampling. We assess robustness to the set of evaluated models by repeatedly fitting the factor model on random subsamples of 80% of models. For each subsample, we compute congruence with the full-data solution. The resulting mean congruence is 0.986 ± 0.028, indicating strong stability under model resampling.

Table 4: Factor congruence across different numbers of factors. High values indicate that the factor structure is stable across dimensionalities.

k_{small}	k_{large}	Mean Congruence	Min Congruence
5	6	0.975	0.918
6	7	0.974	0.896
7	8	0.999	0.999
8	9	0.968	0.752

Consistency stability. We further evaluate the robustness of benchmark-conditioned consistency scores under model subsampling. For each subsample, we recompute $Z(b)$ and examine the resulting benchmark ordering. Benchmark consistency rankings remain unchanged under subsampling, indicating robustness of the observed structure.

Conclusion. Together, these analyses show that the global structure reported in Section 2.2 is robust to normalization choices, factor dimensionality, and model subsampling.

B.5 Tasks with High Uniqueness

We examine the uniqueness terms Ψ from the PAF decomposition to identify tasks not well explained by shared latent factors. Table 5 lists tasks with high uniqueness values (≥ 0.4).

Table 5: Tasks with high uniqueness values. These tasks are less well explained by the shared latent structure.

Task	Uniqueness
MME_OCR	0.47
RealWorldQA	0.40
MMVet_rec,ocr,spat,math	0.89
MMVet_rec,know,spat	0.61
CCBench_historical_figure	0.42

High uniqueness indicates task-specific variability beyond the dominant benchmark-conditioned structure.

B.6 Benchmark Consistency Profiles

We further analyze benchmark-conditioned consistency scores $Z(b)$ introduced in Section 2.2.

Figure 9 shows the baseline consistency scores for each benchmark. We observe substantial variation in internal coherence: some benchmarks (e.g., AI2D, MMBench) exhibit strong consistency across tasks, while others (e.g., MME, MMVet) show weaker structure.

To assess robustness, Figure 10 shows consistency scores under different normalization schemes. The curves largely overlap, and the relative ordering of benchmarks remains unchanged.

These results indicate that benchmark-conditioned consistency is not an artifact of normalization, but reflects intrinsic properties of the evaluation benchmarks.

B.7 Discussion

Across all normalization strategies, we observe:

- Near-perfect stability of factor structure (mean congruence > 0.96)
- Identical ordering of benchmarks by consistency
- Strong preservation of cross-benchmark ranking relationships

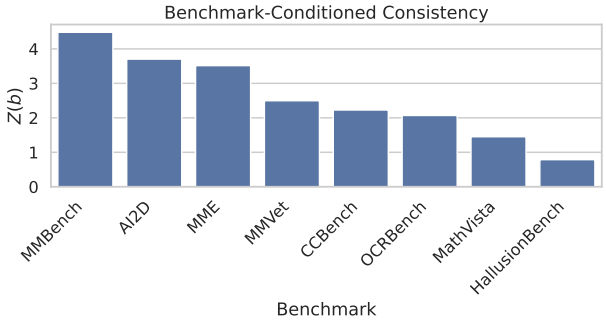


Figure 9: Benchmark-conditioned consistency scores $Z(b)$. Higher values indicate stronger coherence of model behavior within a benchmark.

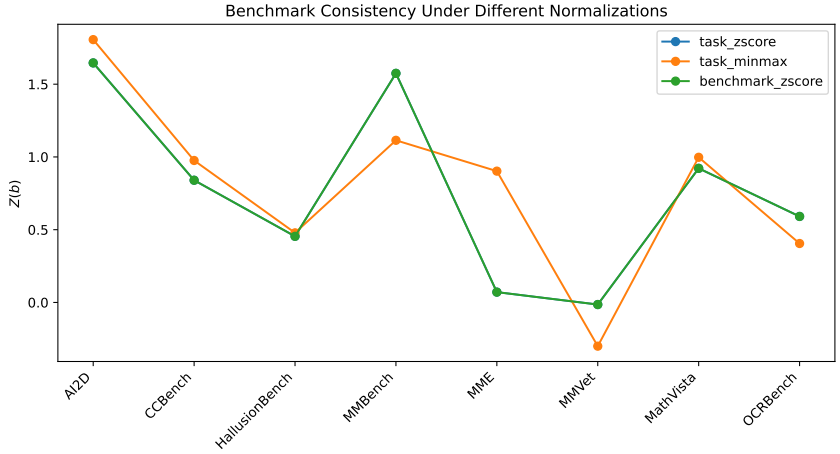


Figure 10: Benchmark consistency scores $Z(b)$ under different normalization schemes. The relative ordering of benchmarks is stable across normalization choices.

These results indicate that the benchmark-conditioned structure identified in Section 2.2 is not an artifact of score scaling, but reflects intrinsic properties of the evaluation benchmarks.

C Additional Benchmark-Level Structure

C.1 Geometric Visualization of Task Space

To provide an intuitive view of the task space, we apply t-SNE van der Maaten & Hinton (2008) to task representations derived from model performance vectors. Figure 11 shows the resulting embedding.

The visualization is consistent with the factor analysis in Section 2.2: tasks from the same benchmark tend to group together, although boundaries are not sharp. This figure is qualitative and is intended as a complement to the quantitative analyses in the main text.

C.2 Benchmark-Level Ranking Agreement

We compute benchmark-induced model rankings by averaging model performance within each benchmark and measuring pairwise agreement using Kendall's τ Kendall (1938). Figure 12 shows the resulting benchmark-level ranking correlation matrix.

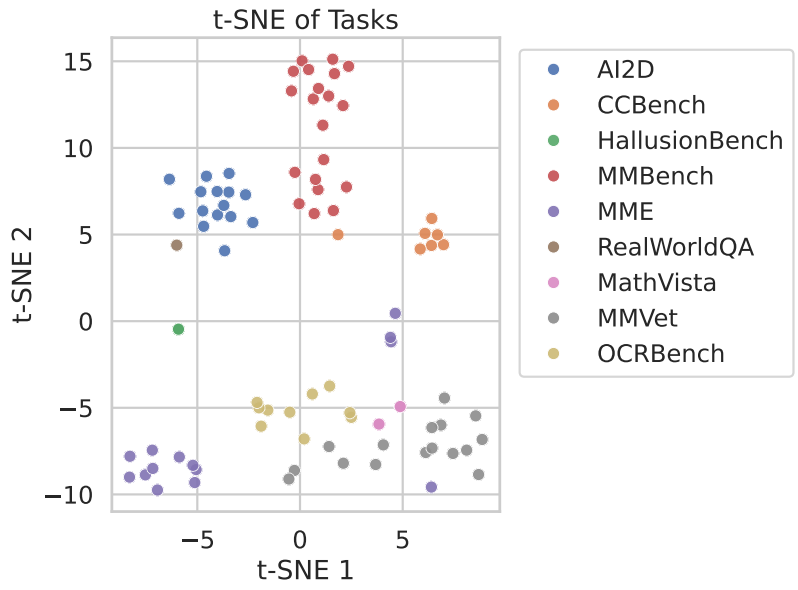


Figure 11: t-SNE embedding of tasks based on model performance. Tasks tend to cluster according to benchmark identity, although clusters are not perfectly separated.

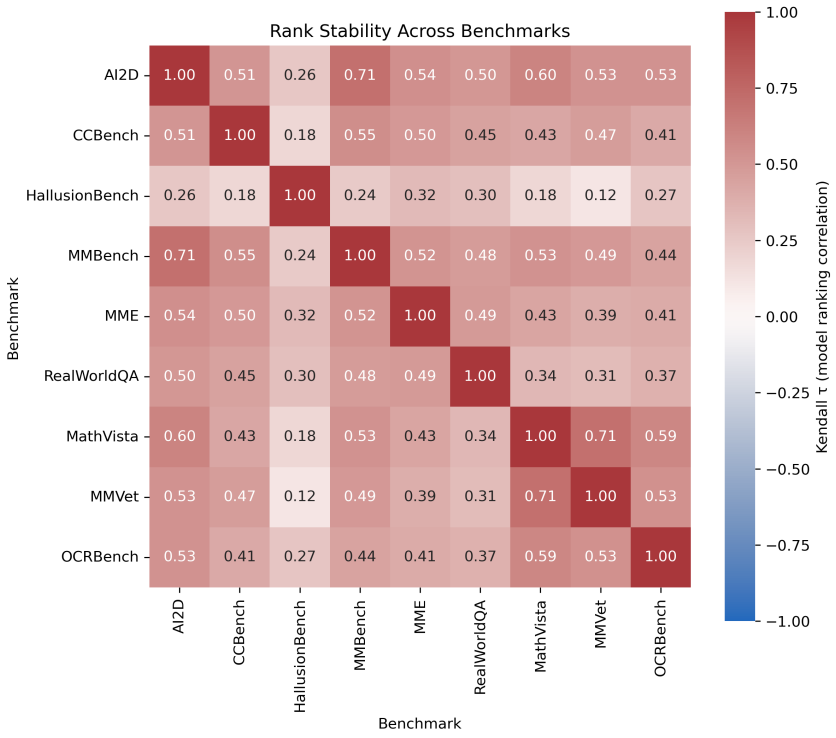


Figure 12: Kendall's τ correlation between benchmark-induced model rankings. Several benchmark pairs exhibit only moderate or low agreement, indicating benchmark-dependent rankings.

Several benchmark pairs exhibit only moderate or low agreement, providing additional evidence that benchmarks are not interchangeable evaluation instruments even when they are used to compare the same set of models.

D Additional Cross-Benchmark Task-Pair Analyses

This section provides additional analyses supporting the cross-benchmark task-pair results in Section 2.2.

D.1 Full Agreement Metrics

Table 6 reports agreement metrics for all evaluated task pairs.

Table 6: Full cross-benchmark agreement metrics.

Task 1	Task 2	Pearson	Spearman	Kendall
MME_celebrity	MMBench_celebrity_recognition	0.49	0.48	0.33
MME_position	MMBench_object_localization	0.66	0.66	0.49
MME_OCR	OCRBench_Regular_Text_Recognition	0.34	0.17	0.11
OCRBench_Regular_Text_Recognition	RealWorldQA	0.38	0.47	0.33

D.2 Rank Agreement Visualization

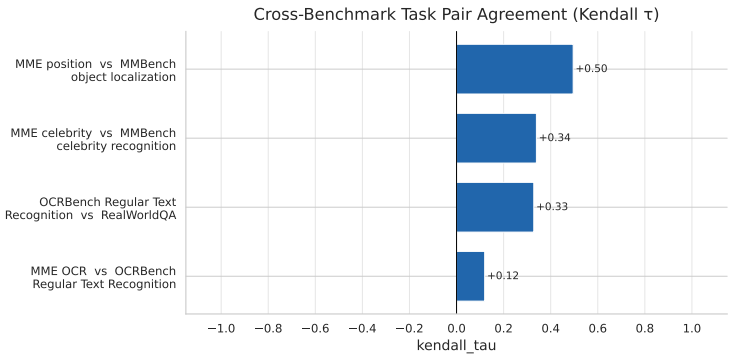


Figure 13: Kendall τ agreement across representative cross-benchmark task pairs. Agreement varies substantially, from weak alignment (e.g., OCR tasks with $\tau \approx 0.12$) to moderate alignment (e.g., localization tasks with $\tau \approx 0.50$).

Figure 13 visualizes ranking agreement across representative task pairs. Agreement varies substantially, from weak alignment to moderate alignment.

Figure 14 shows representative rank agreement patterns.

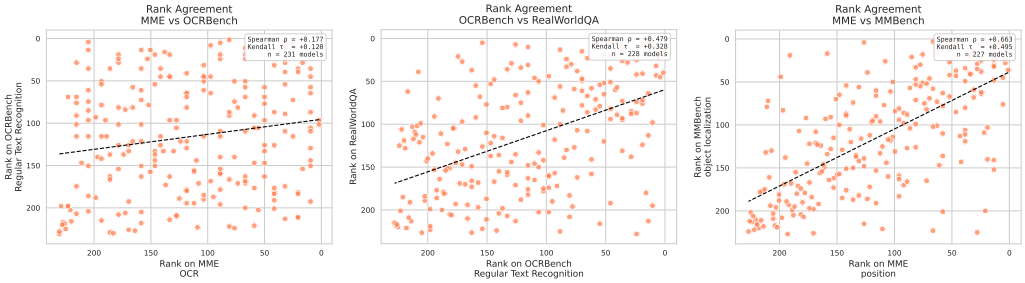


Figure 14: Rank agreement across representative task pairs. Substantial dispersion indicates inconsistent model ordering.

Interpretation. These results reinforce the conclusion that capability labels alone do not determine evaluation behavior: tasks labeled as targeting similar capabilities can induce substantially different model rankings across benchmarks.

D.3 Score-Level Agreement

Figure 15 shows score-level correlations for representative task pairs.

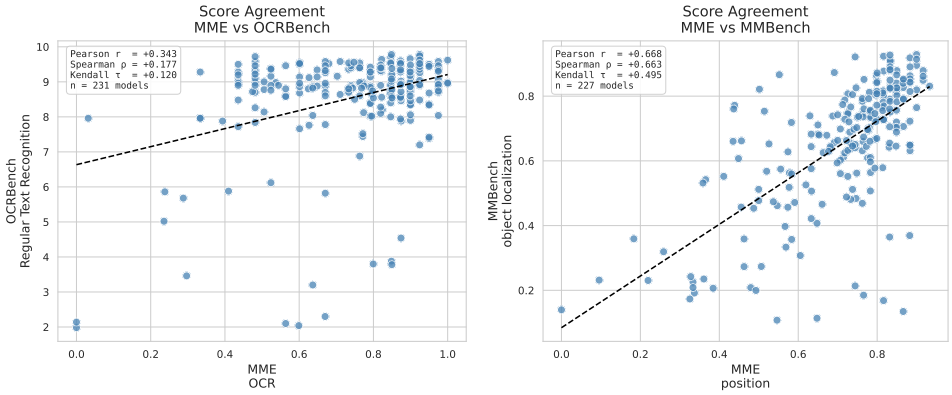


Figure 15: Score-level agreement across task pairs.

D.4 Distribution of Agreement

Figure 16 summarizes the distribution of agreement metrics across evaluated task pairs.

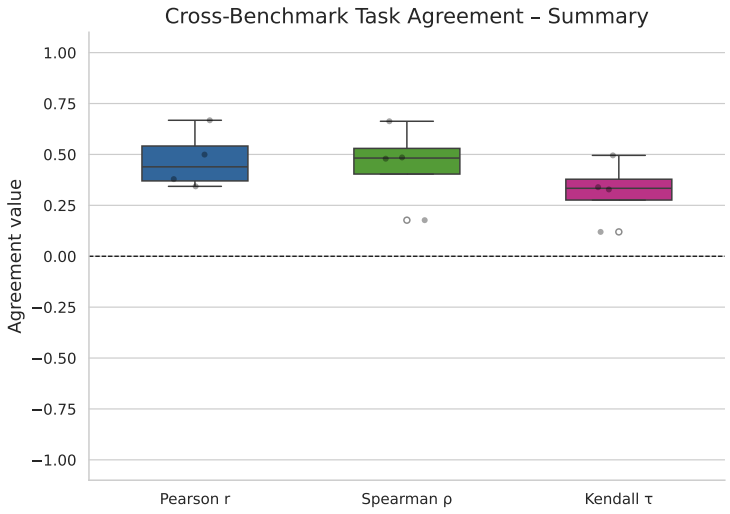


Figure 16: Distribution of agreement across task pairs. Alignment varies substantially across task types.

E Benchmark Augmentation Details

This section describes the construction of the formulation-augmented benchmarks used in Section 3.

E.1 MME

E.1.1 Overview

Starting from the original MME benchmark Fu et al. (2025), which consists of binary yes/no questions, we construct two additional formulations for each example:

- Multiple-choice (MCQ)
- Open-ended (free-form)

All variants share identical visual inputs and semantic targets. The goal is to isolate the effect of task formulation while keeping task content fixed.

E.1.2 Multiple-Choice Construction

For each binary question, we generate a multiple-choice version consisting of one correct answer and four distractors.

Distractor generation. We use 3 LLMs to generate plausible but incorrect answer options conditioned on the original question and image context. Distractors are designed to be semantically related to the correct answer while remaining incorrect. Each of those LLMs independently generated a set of four distractors per item. For each item, we randomly selected one of the three sets, so every MCQ has one correct answer and four distractors from a single generator.

Answer randomization. The correct answer is inserted among the distractors and its position is randomized to avoid positional bias.

Consistency. We manually verify a subset of generated examples to ensure that:

- distractors are not trivially invalid
- the correct answer remains unambiguous

E.1.3 Open-Ended Construction

We convert each binary question into an open-ended prompt using templated transformations.

Prompt reformulation. The original yes/no question is rewritten into a free-form query that elicits a natural language response (e.g., replacing "Is there a cat?" with "What animals are present in the image?").

Semantic preservation. Prompts are constructed to preserve the original semantic target while removing answer constraints. This ensures that differences in performance arise from the response format rather than task reinterpretation.

E.2 Evaluation Protocol

Binary and MCQ. While constrained formats naturally align with exact matching against the ground-truth answer, we subjected these outputs to the same LLM-as-a-judge protocol used for open-ended queries. This unifies the evaluation pipeline and guarantees that the observed performance degradation in open-ended formats is driven by genuine model behavior rather than strictness artifacts of the judging mechanism

Open-ended evaluation. Open-ended responses are evaluated using an LLM-as-a-judge Zheng et al. (2023).

We use `DeepSeek-Reasoner` DeepSeek-AI (2025) to:

- parse model outputs

- compare responses to reference answers
- assign binary correctness labels

The evaluation prompt instructs the judge to determine whether the model's response is semantically correct with respect to the ground truth. To ensure judge reliability and mitigate evaluation noise, all open-ended responses were evaluated three times, with final correctness determined by majority vote.

Human validation of the judge. We sampled 500 model responses, stratified across benchmarks and formulations (weighted toward open-ended responses, where judging is least trivial) and balanced across models. A human annotator, blind to the judge's decision, labeled each response as correct or incorrect against the reference answer. The judge agrees with human labels in 92% of cases (Cohen's κ=83).

Robustness considerations. To reduce evaluation noise:

- responses are normalized prior to evaluation
- ambiguous outputs are resolved using consistent judgment criteria

E.3 Models Evaluated

We evaluate a diverse set of open-source and proprietary MLLMs: Aria, AzureGPT4o-mini, AzureGPT5-4, AzureGPT5-4-mini, Claude4 Sonnet, GLM-4.1V-9B-Thinking, GLM-4.6V-Flash, GeminiFlash2-5, Gemma3-27B, InternVL2-26B, InternVL3.5-30B-A3B, InternVL3.5-38B, Kimi-VL-A3B-Thinking-2506, Ovis2-8B, Qwen2.5-VL-7B-Instruct and molmo-7B-D-0924.

This set includes both open-source and proprietary models spanning a range of architectures and capabilities, including recent multimodal large language models OpenAI (2023); DeepMind (2023); Chen et al. (2023); Bai et al. (2023).

E.4 Discussion

This augmentation procedure enables controlled comparison across task formulations while holding visual input and semantic targets fixed. While LLM-based generation and evaluation may introduce noise, the systematic differences observed in Section 3 persist under exact-match and classification-only analyses, indicating that they are not driven by judge-based scoring.

F Ranking Stability Across Task Formulations

This appendix provides additional evidence for the instability of model rankings across task formulations.

F.1 Pairwise Rank Agreement

Figure 17 visualizes pairwise model rankings across Binary, MCQ, and Open-ended formulations for MME.

Figure 17 shows substantial dispersion across all formulation pairs. Even when formulations share identical images and semantic goals, model rankings are not preserved.

This result complements Table 2 in the main paper and provides a visual illustration of formulation-induced ranking instability.

F.2 Within-Format Baseline and Confidence Intervals

Motivation. Low cross-format agreement is meaningful only relative to how stable a ranking is when the formulation does *not* change. With a finite number of items and closely matched models, two rankings computed on different items of the same formulation will not agree perfectly. We therefore compare cross-format agreement against a within-format baseline that captures this item-sampling noise.

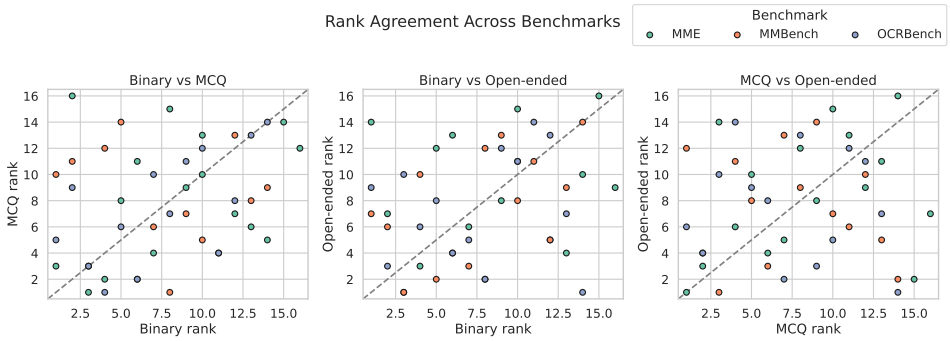


Figure 17: Rank agreement across formulations. Each point is a model; the dashed line marks identical ranks. Points far from the diagonal are models whose rank changes with the formulation.

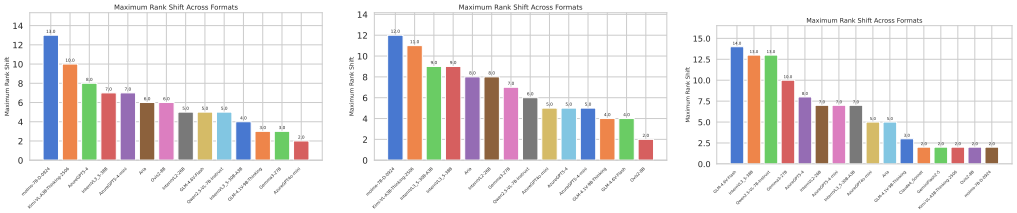


Figure 18: Maximum rank shift across formulations for each model.

Procedure. In each of 1,000 bootstrap iterations, we draw two disjoint samples A and B of 200 items each (without replacement) from a benchmark's item pool. For each formulation f, the within-format agreement is $\tau(f_A, f_B)$, the Kendall τ between model rankings computed on A and on B under the same formulation. For each formulation pair (f, g), the cross-format agreement is $\tau(f_A, g_A)$, computed on the *same* items. This comparison is conservative: cross-format rankings share items, which favors agreement, whereas within-format rankings do not. The *Within-format* row of Table 2 averages $\tau(f_A, f_B)$ over formulations. All $\pm$ values are half-widths of 95% percentile intervals over iterations. Because subsamples contain 200 items rather than the full pool, these intervals are conservative.

Significance. For each pair (f, g) and iteration, we compute

$$\Delta_{fg} = \tfrac{1}{2}\big(\tau(f_A, f_B) + \tau(g_A, g_B)\big) - \tau(f_A, g_A).$$

A pair is marked * when the 2.5th percentile of Δ_{fg} is above zero, i.e., when the drop in agreement caused by changing the formulation exceeds what re-sampling items alone produces in at least 97.5% of iterations.

Interpretation. High within-format agreement indicates that a formulation ranks models reliably; cross-format agreement significantly below it indicates that the formulation itself, rather than item-sampling noise, changes model rankings. Within-format agreement is high on MME and MMBench. On OCRBench, within-format agreement for Binary is lower (τ=0.50) because most models score near ceiling, leaving little room to separate them; low cross-format agreement involving Binary on OCRBench is therefore partly attributable to this ceiling effect.

F.3 Rank Shift Across Formulations

To further quantify ranking instability, we measure the maximum rank shift for each model:

$$\text{Shift}(m) = \max_{f_1, f_2} |r_m^{(f_1)} - r_m^{(f_2)}|.$$

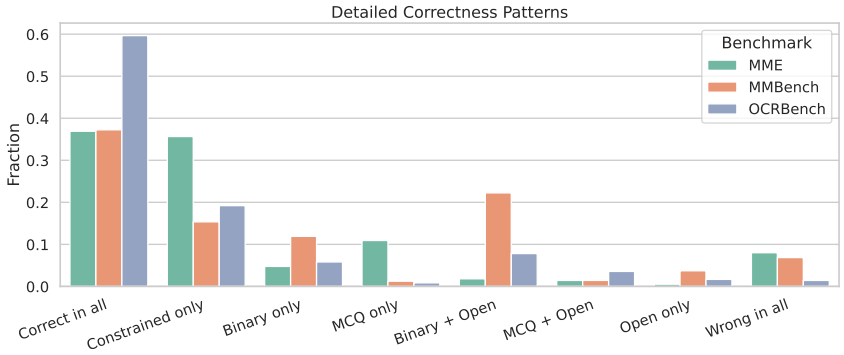


Figure 19: detailed correctness-pattern distribution across benchmarks.

Figure 18 shows that formulation effects are not uniform across models. While some models remain relatively stable, others undergo substantial ranking changes depending on how the task is formulated.

Interpretation. These results confirm that task formulation affects not only absolute performance but also relative model comparison. Ranking instability therefore reflects a systematic consequence of evaluation design rather than random evaluation noise.

G Instance-Level Analysis Across Benchmarks

This appendix provides detailed analyses supporting the instance-level findings in Section 3.

G.1 Unified Correctness Patterns

Figure 19 summarizes correctness-pattern distributions across MME, MMBench, and OCRBench.

Across all three benchmarks, the dominant formulation-sensitive pattern corresponds to constrained-only correctness: examples that are solved under Binary and MCQ formulations but fail under Open-ended evaluation.

Open-only patterns remain comparatively rare.

Interpretation. The same asymmetry appears across benchmark families despite substantial differences in task content and benchmark design. This indicates that formulation sensitivity is not specific to a particular benchmark.

G.2 Detailed Correctness Patterns

Figure 3(c) reports the full distribution over the eight possible correctness combinations.

While several mixed patterns exist, constrained-only correctness remains the dominant formulation-sensitive behavior.

G.3 Cross-Formulation Agreement

Table 7 reports agreement between formulations at the instance level.

Agreement is substantially higher between constrained formulations than between constrained and Open-ended evaluation.

This confirms that task formulation changes which examples are solved correctly rather than merely rescaling overall performance.

Table 7: Instance-level agreement across formulations. Agreement denotes the fraction of model–item pairs receiving the same correctness label under both formulations, while Jaccard measures overlap among correctly solved examples. Across all benchmarks, constrained formulations exhibit substantially higher agreement with each other than with Open-ended evaluation.

Benchmark	Formulation Pair	Agreement	Jaccard
MME	Binary vs. MCQ	High	High
	Binary vs. Open	Moderate	Low
	MCQ vs. Open	Moderate	Low
MMBench	Binary vs. MCQ	Moderate	Moderate
	Binary vs. Open	Moderate	Moderate
	MCQ vs. Open	Moderate	Moderate
OCRBench	Binary vs. MCQ	High	High
	Binary vs. Open	Moderate	Moderate
	MCQ vs. Open	High	Moderate

G.4 Aggregate Performance Across Formulations

Figure 3 (a) reports average performance under Binary, MCQ, and Open-ended formulations for all three benchmarks.

Across all benchmarks, constrained formulations consistently yield higher performance than Open-ended evaluation.

Although the absolute performance levels differ between benchmarks, the ordering of formulations remains remarkably stable: Binary and MCQ outperform Open-ended evaluation in every benchmark.

Interpretation. The consistency of this pattern suggests that formulation effects are not tied to a specific benchmark design. Instead, constrained formulations appear to systematically simplify the prediction problem by reducing the effective answer space.

H Formulation Sensitivity Across Categories and Benchmarks

We next investigate whether formulation sensitivity is restricted to specific task categories.

H.1 Category-Level Correctness Patterns

Figure 20 shows that formulation-sensitive behavior appears across diverse task categories.

H.2 Category-Level Summary Statistics

Table 8 reports the proportion of examples exhibiting constrained-only and open-only correctness patterns.

Across most categories, constrained-only correctness substantially exceeds open-only correctness.

This indicates that the asymmetry between constrained and open-ended evaluation is not tied to a small number of categories but appears broadly across task types.

H.3 Model-Level Instability

Figure 21 shows that formulation sensitivity is distributed across models rather than concentrated in a small subset of systems.

H.4 Aggregated Category Performance

Figure 22 shows that constrained formulations consistently yield higher performance than Open-ended evaluation across categories.

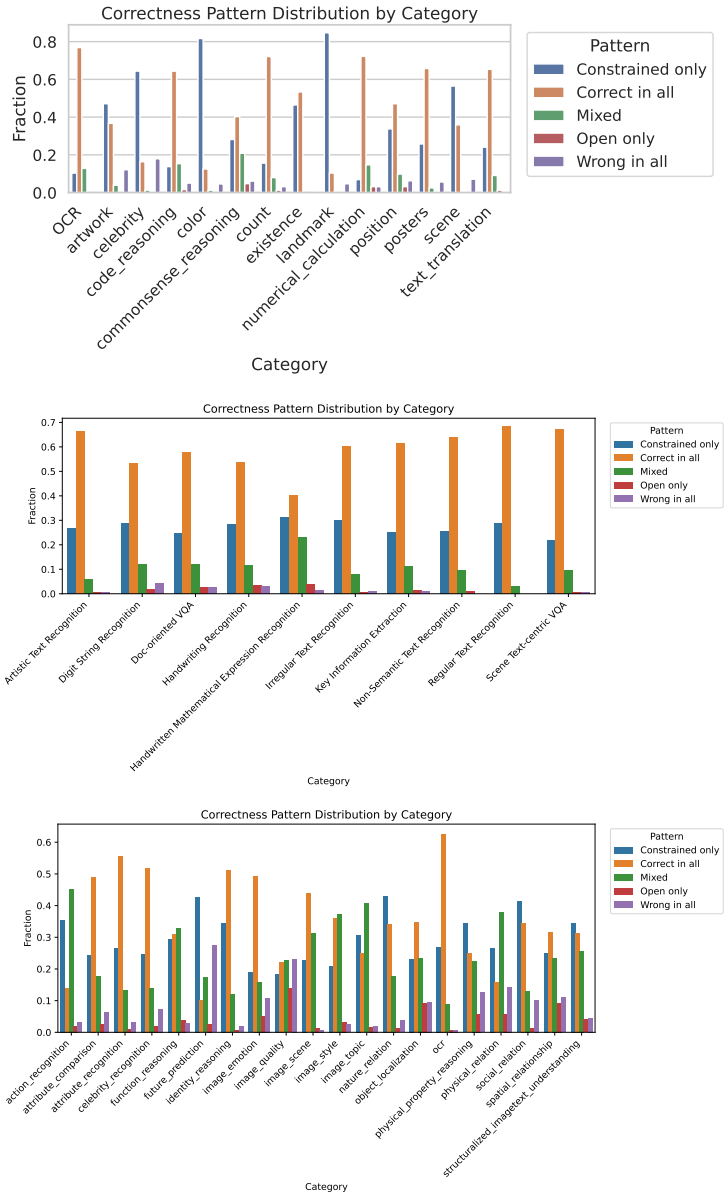


Figure 20: Correctness-pattern distributions across different categories.

Summary. Across categories, models, and items, constrained formulations systematically increase success rates relative to Open-ended evaluation. This provides a mechanistic explanation for the aggregate ranking instability observed in Section 3.

This section provides additional analysis of how task formulation affects consistency at the task level.

H.5 Task-Level Consistency Across Formats

We measure consistency for each task by computing agreement across its format variants. Table 9 reports the mean within-task correlation and the corresponding standardized consistency score.

Interpretation. Consistency varies substantially across tasks. Some tasks remain stable across formats, while others exhibit strong format sensitivity, with model behavior changing significantly under different interaction regimes.

Table 8: Fraction of examples per category (MME) exhibiting format-sensitive correctness patterns.

Category	Constrained-only (%)	Open-only (%)
OCR	8.95	0.53
Artwork	41.92	0.55
Celebrity	56.53	0.15
Code Reasoning	10.79	2.63
Color	78.07	0.00
Common_sense_reasoning	25.86	6.02
Counting	12.46	3.51
Existence	46.84	0.00
Landmark	80.95	0.00
numerical_calculation	5.00	5.00
Position	28.95	5.96
Posters	22.34	0.47
Scene	54.37	0.05
Text Recognition	19.21	1.05

Table 9: Task-level consistency across formats for MME benchmark. Positive values indicate stable behavior across formats, while low or negative values indicate strong format sensitivity.

Task	Mean Within Corr	Z-Consistency
Posters	0.77	1.69
Text Translation	0.70	1.50
Code Reasoning	0.64	1.33
Numerical Calculation	0.64	1.32
Celebrity	0.63	1.29
Count	0.50	0.92
Commonsense Reasoning	0.38	0.58
Landmark	0.37	0.56
Artwork	0.29	0.32
Color	0.26	0.24
Existence	0.17	-0.01
Position	0.13	-0.11
OCR	0.08	-0.27
Scene	-0.12	-0.84

This supports the interpretation in Section 3 that evaluation instability is not uniform, but driven by specific task types.

I Formulation-Conditioned Structure

In Section 3, we showed that changing task formulation affects aggregate scores, model rankings, and instance-level correctness. Here we provide additional structural analyses showing that format also contributes to the covariance structure of model behavior.

I.1 Setup

We construct a model-by-task matrix over the augmented MME dataset, where each column corresponds to a category–format pair, such as OCR–Binary, OCR–MCQ, or OCR–Open-ended. Since all formats are derived from the same underlying examples, variation across columns reflects category and task formulation rather than different image content.

I.2 Cross-Format Ranking Agreement

Figure **??** reports model-ranking agreement across formats using Pearson, Spearman, and Kendall correlations. The figure shows that Binary and MCQ rankings are more closely aligned than either is with Open-ended evaluation. This supports the main finding that task formulation affects model ordering. Agreement is consistently lowest between the most dissimilar formulations, Binary and Open-ended ($\tau \leq 0.10$ on all three benchmarks), while the most aligned pair varies by benchmark.

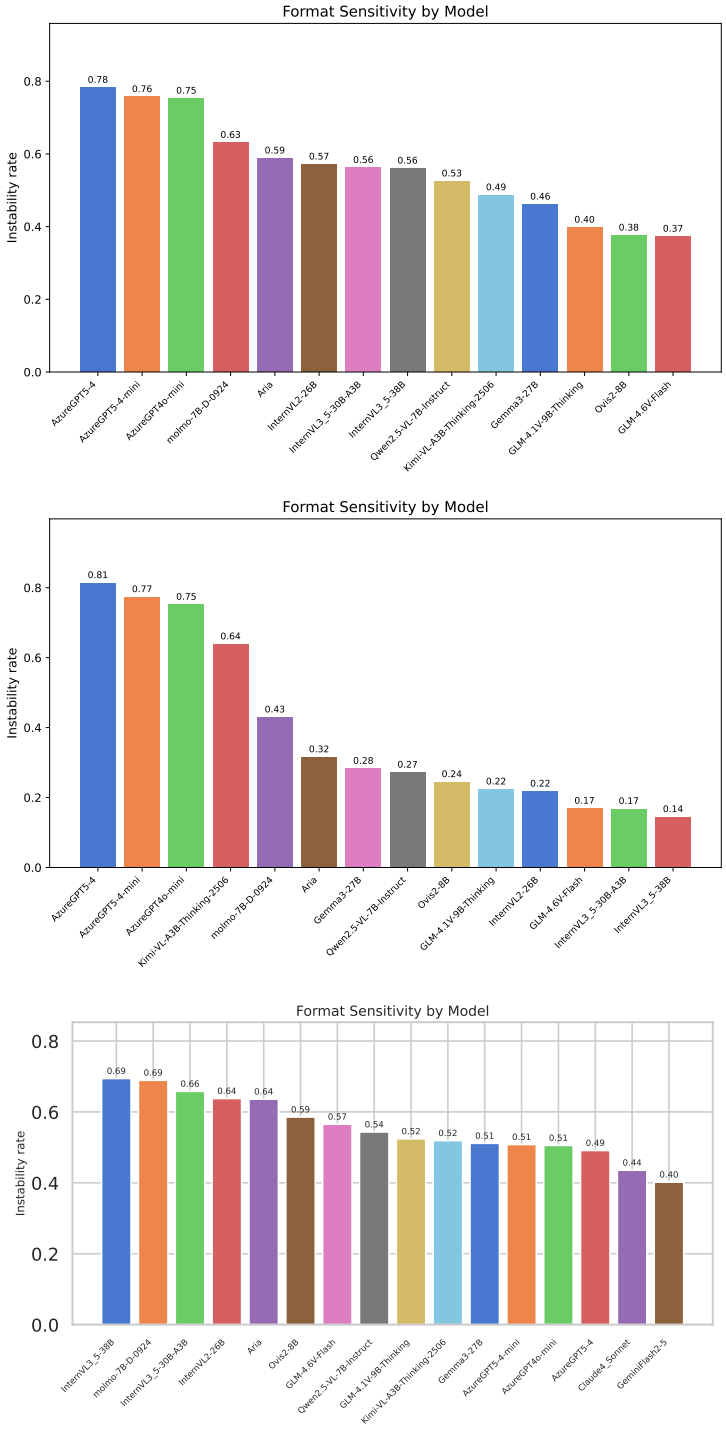


Figure 21: Format sensitivity per model: fraction of items whose correctness differs across Binary, MCQ, and Open-ended formulations, for MME (top), MMBench (middle), and OCRBench (bottom). [Claude-4 Sonnet and Gemini-2.5 Flash were evaluated on OCRBench only.]

At the instance level, constrained formulations generally agree more with each other than with Open-ended evaluation (Table **??**). Formulation therefore affects both which items are solved and how models are ordered, and its effect on rankings is not predictable from the benchmark alone.

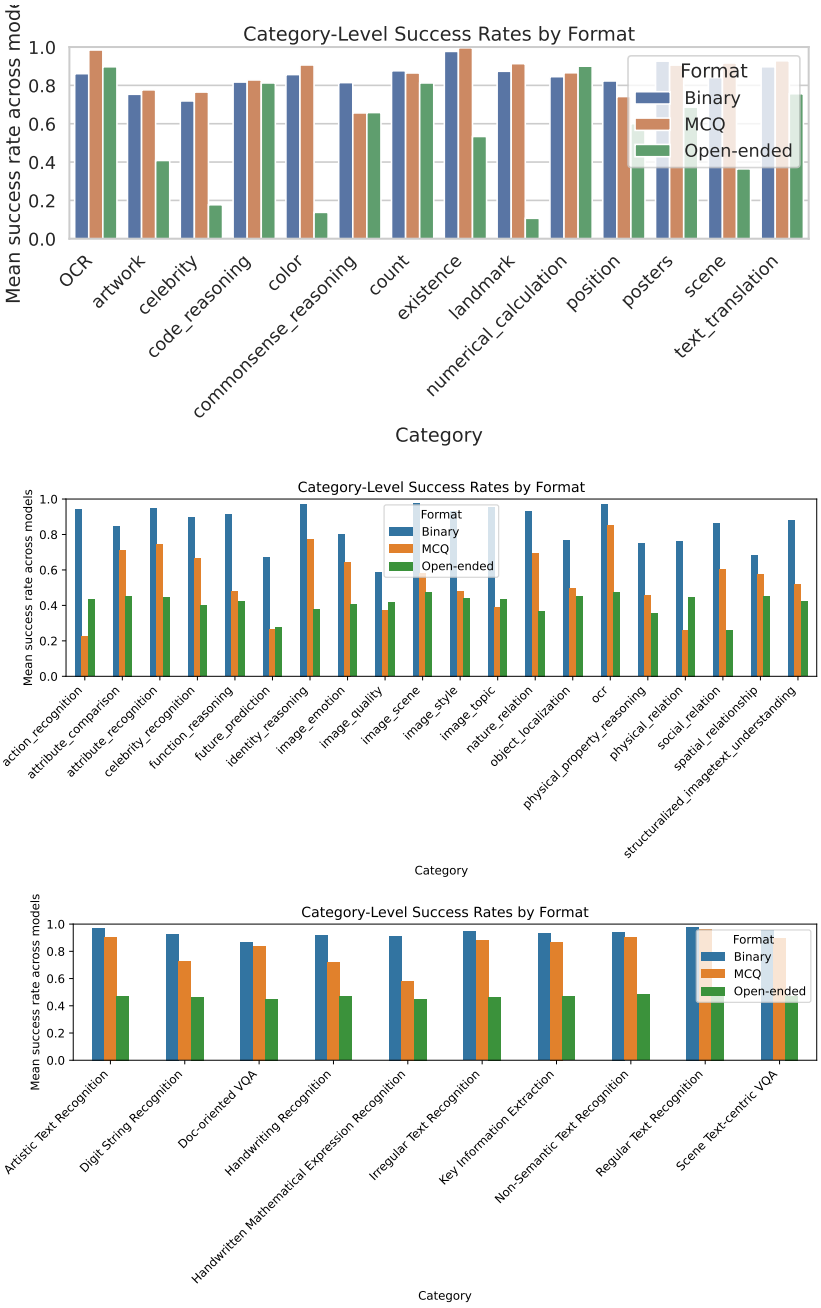


Figure 22: Mean success rate per category and formulation.

I.3 Within- vs. Cross-Format Agreement

Figure 23 compares agreement within formats, across formats for the same task, and across both tasks and formats.

Figure 23 shows that cross-format agreement for the same task is substantially lower than within-format agreement. This indicates that changing format can alter model behavior even without changing task content.

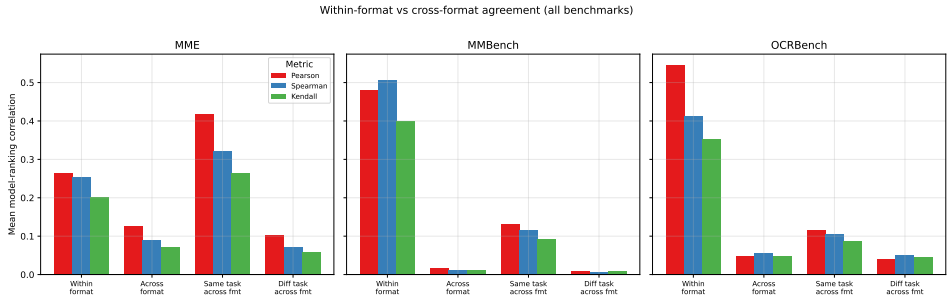


Figure 23: Mean correlation under different comparison conditions. Agreement is highest within the same format and lower across formats, even when the underlying task is fixed.

I.4 Task Correlation Structure

Figure 24 shows the task–task correlation matrix ordered by format.

Figure 24 shows partial grouping by task formulation. Open-ended variants display distinct correlation patterns relative to constrained formats, suggesting that format contributes to the latent structure of model behavior.

I.5 Factor Stability and Base Task Structure

Figure 25 reports factor congruence across different numbers of factors.

Figure 26 shows model-ranking agreement across base tasks after ignoring format.

Figures 25 and 26 show that the augmented MME structure is not explained by format alone. Task/category structure also contributes. This supports the interpretation that format is one important source of evaluation variability, rather than the sole cause.

Summary. The augmented MME analysis shows that task formulation contributes to the structure of model behavior: ranking agreement is lowest between the most dissimilar formulations, and cross-format agreement for the same task is lower than within-format agreement (Figure **??**). This provides structural support for the controlled findings in Section 3.

J Item Response Theory for Subset Selection

We use Item Response Theory (IRT) (Baker, 2001; Lord, 1980) as a principled framework for selecting informative evaluation items.

In the standard 2PL IRT model, the probability that model m correctly answers item i is:

$$P(c_{m,i} = 1) = \sigma(a_i(\theta_m - b_i)),$$

where θ_m is model ability, b_i is item difficulty, and a_i is the discrimination parameter.

Standard IRT selection. Standard IRT assumes a single evaluation regime and estimates one difficulty parameter per item. This captures dominant average structure but does not model formulation-dependent variation.

Formulation-aware IRT. To account for task formulation, we allow item difficulty to vary across formulations:

$$P(c^{(f)}_{m,i} = 1) = \sigma(a_i(\theta_m - b_{i,f})),$$

where f indexes formulation type. This allows the same underlying item to have different effective difficulty under Binary, MCQ, and Open-ended formulations.

This formulation captures the empirical observation in Section 3: the same image and semantic goal may be easy under one formulation and difficult under another.

Selection. For each item and formulation, Fisher information is $I_{i,f}(\theta) = a_i^2 \, P_{i,f}(\theta) \, (1 - P_{i,f}(\theta))$. We score each item by $\bar{I}_i = \frac{1}{|\mathcal{F}| \, |\mathcal{M}|} \sum_f \sum_m I_{i,f}(\hat{\theta}_m)$ and select the top-k items. Informative items therefore combine high discrimination with difficulty near the abilities of the evaluated models.

Comparison to standard IRT. Figure 27 compares standard and formulation-aware IRT across the three benchmarks. Formulation-aware IRT consistently improves worst-case ranking fidelity. The improvement is especially pronounced in benchmarks where standard IRT selects subsets that preserve average rankings but reverse under a specific formulation.

Interpretation. Standard IRT captures average item informativeness but ignores evaluation variability induced by task formulation. Formulation-aware IRT explicitly models this variability, producing subsets that better preserve robustness under formulation shifts.

Generalization to held-out models. To test whether selected subsets transfer to models not used for selection, we partition the models into five folds of three; the remaining model is always in the training set. In each fold, IRT parameters are estimated and the subset is selected using the 13 training models only. All 16 models are then ranked by their accuracy on the selected subset.

For each held-out model m, we compute the *rank error* $|r_m^{\text{full}} - r_m^{\text{sub}}|$, where r_m^{full} is its rank among all 16 models on the full data and r_m^{sub} is its rank on the subset. For example, a model ranked 4th on the full data and 5th on the subset has rank error 1. We average over the 15 held-out cases (5 folds $\times$ 3 models).

We compare this against two references. First, the rank error of the *training* models, which reflects how accurately the subset ranks models even when they were used for selection; since the subset is much smaller than the full data and some models have nearly identical scores, this error is not zero. Second, the expected error under random placement: a model assigned a uniformly random rank among n models deviates from its true rank by $(n^2 - 1)/(3n)$ positions on average, which is 5.3 for n=16. Held-out error close to training error indicates that the subset generalizes to unseen models.

K Subset Size vs. Evaluation Metrics

This section provides additional analysis of how subset size affects ranking and score fidelity.

Figure 28 shows that score fidelity improves gradually with subset size across all three benchmarks. In contrast to ranking fidelity, which saturates quickly, accurate estimation of absolute performance requires substantially more samples.

The formulation-augmented benchmarks generally exhibit higher MAE than the original benchmarks, reflecting the additional variability introduced by evaluating multiple task formulations.

Interpretation. Ranking structure is often determined by large relative differences between models and can therefore be recovered with relatively few informative items. Absolute score estimation depends on finer-grained variation across many examples and formulations, making it more sample-intensive.

L Category-Level Subset Diagnostics

This section analyzes how subset selection interacts with semantic task categories.

L.1 Category Bias Across Benchmarks

Figure 29 shows that selected subsets do not necessarily preserve the category distribution of the full benchmark. Category bias decreases with subset size but remains nonzero across benchmarks.

This indicates that IRT-based selection prioritizes discriminative items rather than semantic representativeness.

L.2 MME Category Diagnostics

Figures 30–32 show that selected subsets may be highly concentrated in a small number of categories. However, high category representation does not necessarily imply lower error, and underrepresented categories do not always exhibit the largest MAE.

L.3 MMBench Category Diagnostics

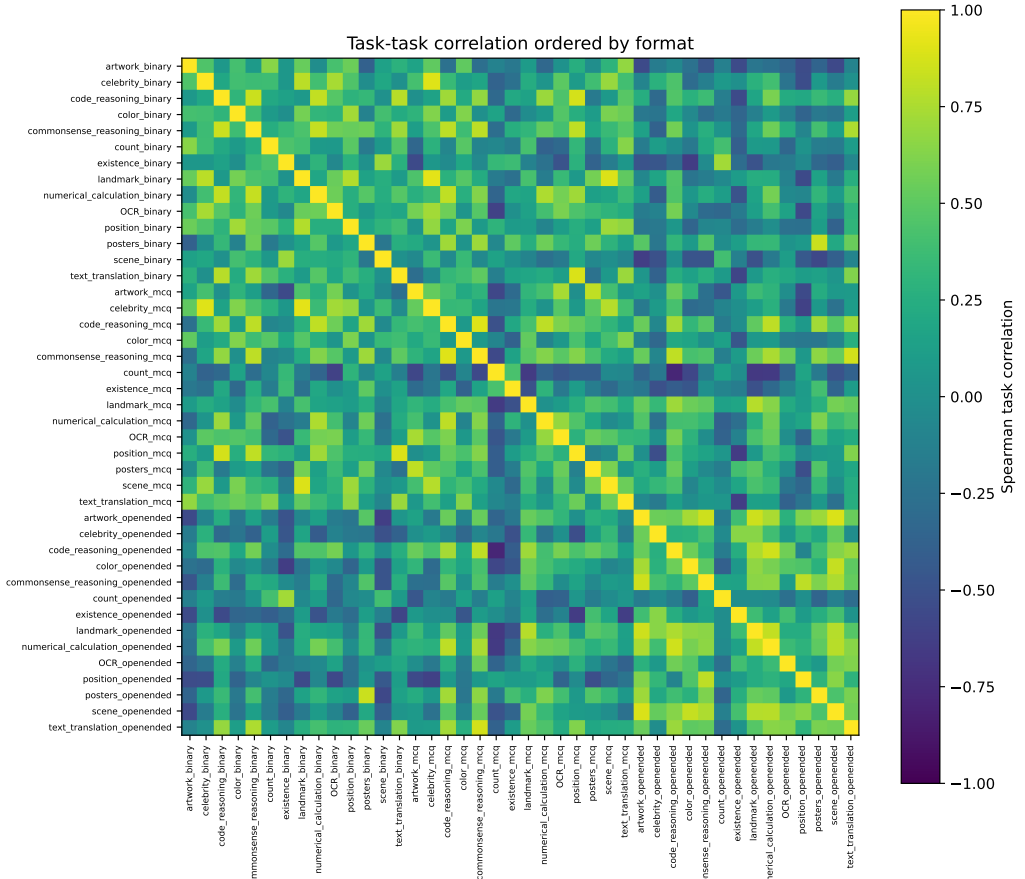
Figures 33–35 show that formulation-aware selection again produces category-skewed subsets. The selected distribution emphasizes categories that are informative for distinguishing models, rather than matching the full benchmark distribution.

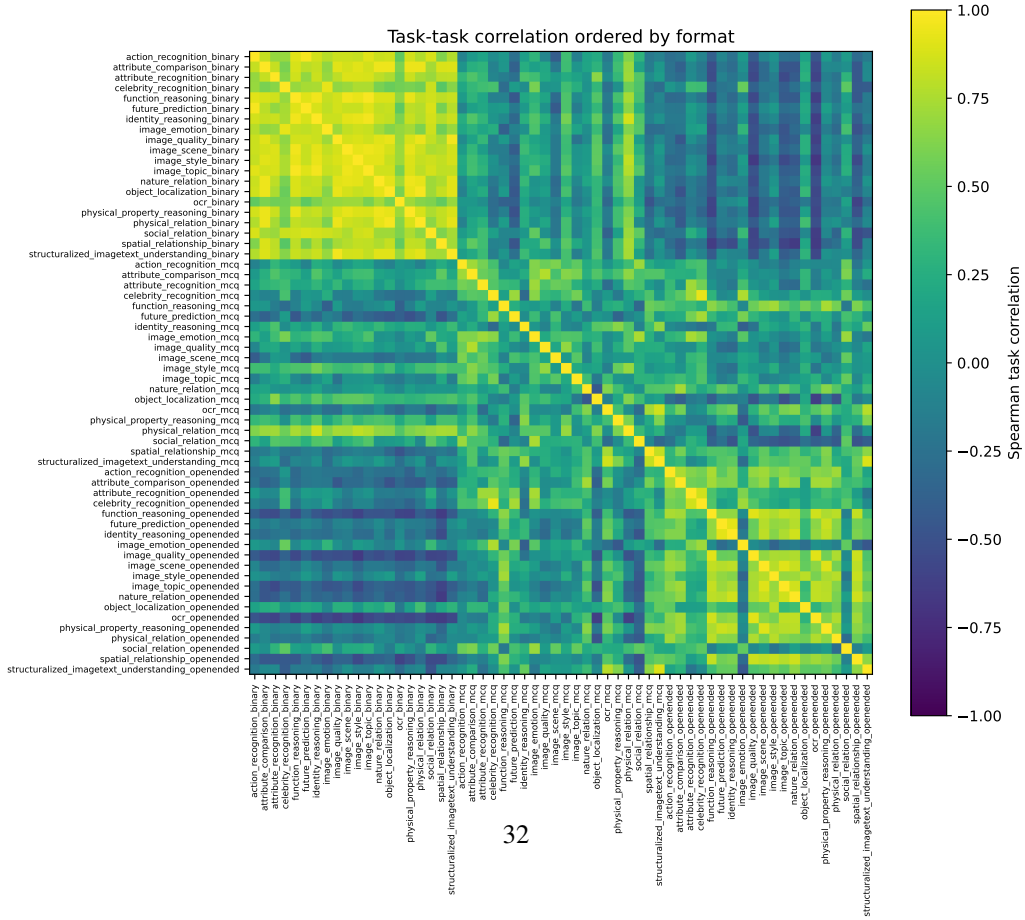
L.4 OCRBench Category Diagnostics

Figures 36–38 show that category-level error varies substantially across OCRBench categories. As in MME and MMBench, preserving the full category distribution is not the objective of IRT-based selection.

Interpretation. Across benchmarks, formulation-aware subset selection prioritizes latent discriminative structure rather than semantic category balance. This produces subsets that may be category-skewed but remain effective for preserving model rankings.

Summary. Category diagnostics show that subset selection is not equivalent to stratified sampling. Informative subsets can deviate substantially from the original category distribution while still preserving ranking fidelity. This supports the broader view that reliable compressed evaluation depends on capturing model-behavior structure, not merely preserving semantic coverage.


Task-task correlation ordered by format
artwork_binary
celebrity_binary
code_reasoning_binary
color_binary
commonsense_reasoning_binary
count_binary
existence_binary
landmark_binary
numerical_calculation_binary
OCR_binary
position_binary
posters_binary
scene_binary
text_translation_binary
artwork_mcq
celebrity_mcq
code_reasoning_mcq
color_mcq
commonsense_reasoning_mcq
count_mcq
existence_mcq
landmark_mcq
numerical_calculation_mcq
OCR_mcq
position_mcq
posters_mcq
scene_mcq
text_translation_mcq
artwork_openended
celebrity_openended
code_reasoning_openended
color_openended
commonsense_reasoning_openended
count_openended
existence_openended
landmark_openended
numerical_calculation_openended
OCR_openended
position_openended
posters_openended
scene_openended
text_translation_openended
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Spearman task correlation


Task-task correlation ordered by format
action_recognition_binary
attribute_comparison_binary
attribute_recognition_binary
celebrity_recognition_binary
function_reasoning_binary
future_prediction_binary
identity_reasoning_binary
image_emotion_binary
image_quality_binary
image_scene_binary
image_style_binary
image_topic_binary
nature_relation_binary
object_localization_binary
ocr_binary
physical_property_reasoning_binary
physical_relation_binary
social_relation_binary
spatial_relationship_binary
structuralized_imagetext_understanding_binary
action_recognition_mcq
attribute_comparison_mcq
attribute_recognition_mcq
celebrity_recognition_mcq
function_reasoning_mcq
future_prediction_mcq
identity_reasoning_mcq
image_emotion_mcq
image_quality_mcq
image_scene_mcq
image_style_mcq
image_topic_mcq
nature_relation_mcq
object_localization_mcq
ocr_mcq
physical_property_reasoning_mcq
physical_relation_mcq
social_relation_mcq
spatial_relationship_mcq
structuralized_imagetext_understanding_mcq
action_recognition_openended
attribute_comparison_openended
attribute_recognition_openended
celebrity_recognition_openended
function_reasoning_openended
future_prediction_openended
identity_reasoning_openended
image_emotion_openended
image_quality_openended
image_scene_openended
image_style_openended
image_topic_openended
nature_relation_openended
object_localization_openended
ocr_openended
physical_property_reasoning_openended
physical_relation_openended
social_relation_openended
spatial_relationship_openended
structuralized_imagetext_understanding_openended
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Spearman task correlation

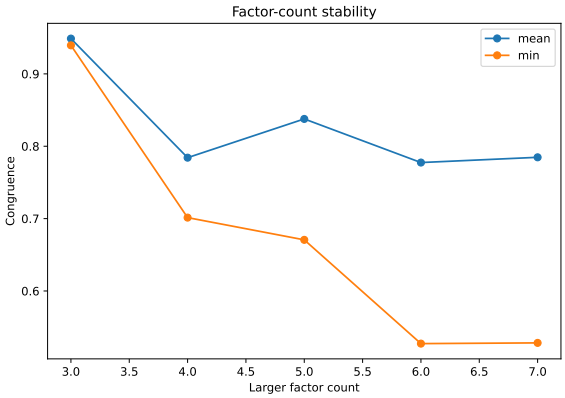
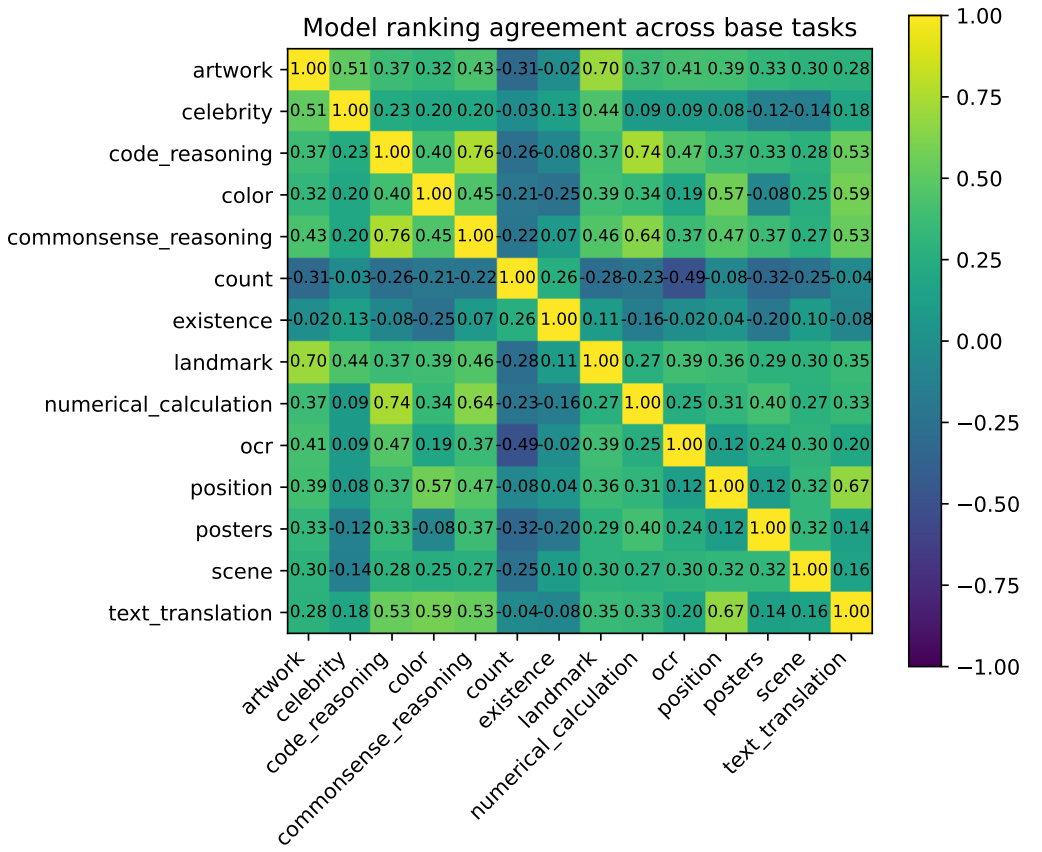
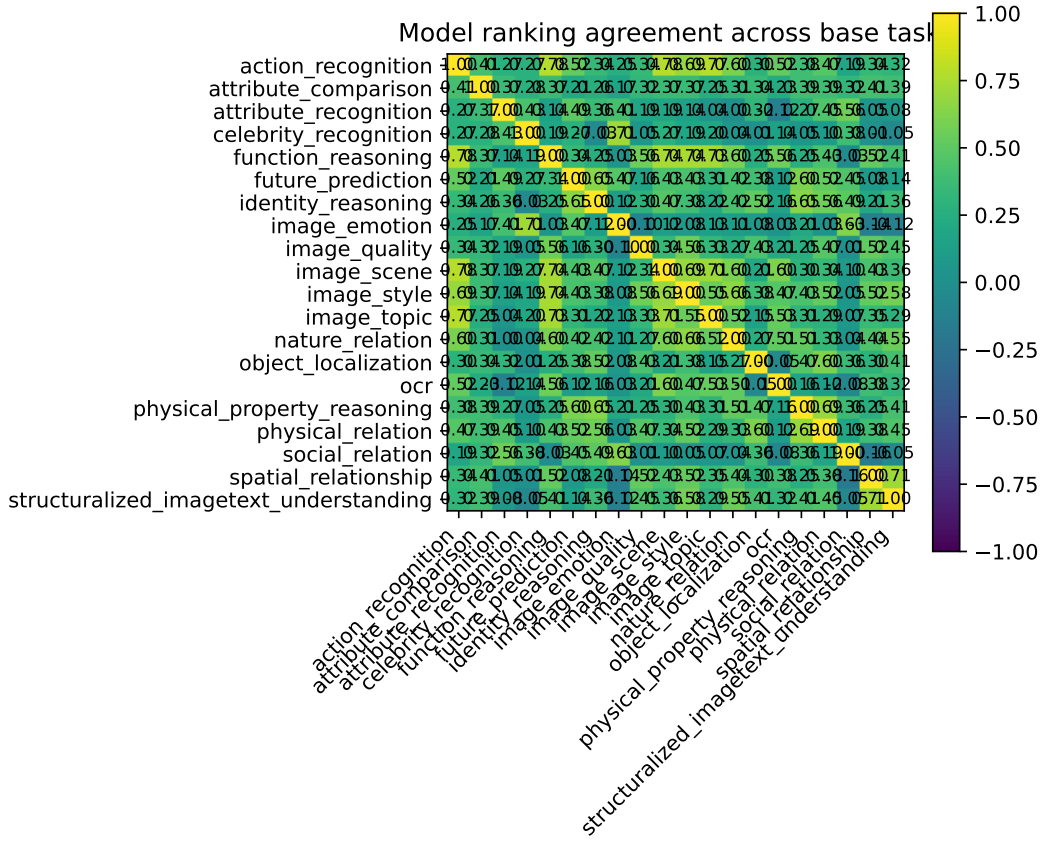


Figure 25: Factor congruence across different numbers of factors. The structure is moderately stable across dimensionalities.


Model ranking agreement across base tasks
artwork 1.00 0.51 0.37 0.32 0.43 -0.31 -0.02 0.70 0.37 0.41 0.39 0.33 0.30 0.28
celebrity 0.51 1.00 0.23 0.20 0.20 -0.03 0.13 0.44 0.09 0.09 0.08 -0.12 -0.14 0.18
code_reasoning 0.37 0.23 1.00 0.40 0.76 -0.26 -0.08 0.37 0.74 0.47 0.37 0.33 0.28 0.53
color 0.32 0.20 0.40 1.00 0.45 -0.21 -0.25 0.39 0.34 0.19 0.57 -0.08 0.25 0.59
commonsense_reasoning 0.43 0.20 0.76 0.45 1.00 -0.22 0.07 0.46 0.64 0.37 0.47 0.37 0.27 0.53
count -0.31 -0.03 -0.26 -0.21 -0.22 1.00 0.26 -0.28 -0.23 -0.49 -0.08 -0.32 -0.25 -0.04
existence -0.02 0.13 -0.08 -0.25 0.07 0.26 1.00 0.11 -0.16 -0.02 0.04 -0.20 0.10 -0.08
landmark 0.70 0.44 0.37 0.39 0.46 -0.28 0.11 1.00 0.27 0.39 0.36 0.29 0.30 0.35
numerical_calculation 0.37 0.09 0.74 0.34 0.64 -0.23 -0.16 0.27 1.00 0.25 0.31 0.40 0.27 0.33
ocr 0.41 0.09 0.47 0.19 0.37 -0.49 -0.02 0.39 0.25 1.00 0.12 0.24 0.30 0.20
position 0.39 0.08 0.37 0.57 0.47 -0.08 0.04 0.36 0.31 0.12 1.00 0.12 0.32 0.67
posters 0.33 -0.12 0.33 -0.08 0.37 -0.32 -0.20 0.29 0.40 0.24 0.12 1.00 0.32 0.14
scene 0.30 -0.14 0.28 0.25 0.27 -0.25 0.10 0.30 0.27 0.30 0.32 0.32 1.00 0.16
text_translation 0.28 0.18 0.53 0.59 0.53 -0.04 -0.08 0.35 0.33 0.20 0.67 0.14 0.16 1.00
artwork
celebrity
code_reasoning
color
commonsense_reasoning
count
existence
landmark
numerical_calculation
ocr
position
posters
scene
text_translation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Model ranking agreement across base task
action_recognition
attribute_comparison
attribute_recognition
celebrity_recognition
function_reasoning
future_prediction
identity_reasoning
image_emotion
image_quality
image_scene
image_style
image_topic
nature_relation
object_localization
ocr
physical_property_reasoning
physical_relation
social_relation
spatial_relationship
structuralized_imagetext_understanding
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Model ranking agreement across base
1.00

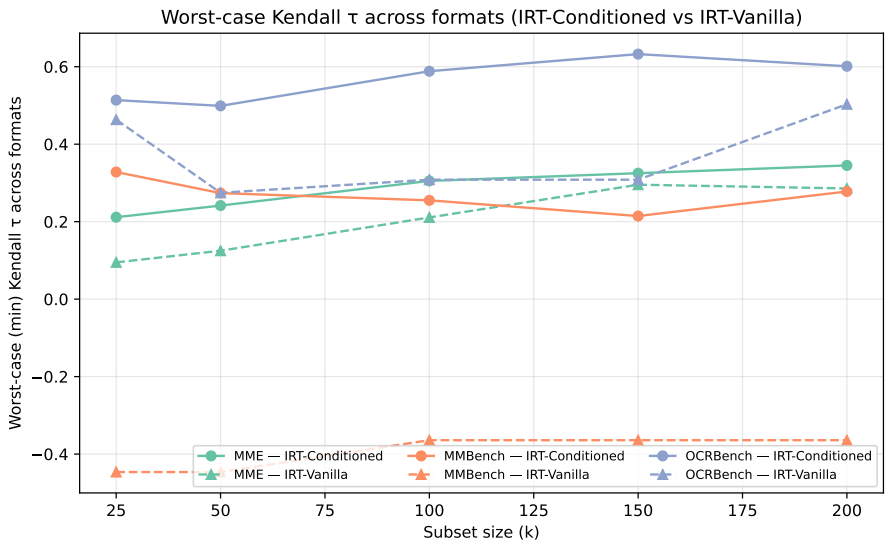


Figure 27: Worst-case Kendall $\tau_{\min}$ across formulations for standard IRT and formulation-aware IRT across MME, MMBench, and OCRBench.

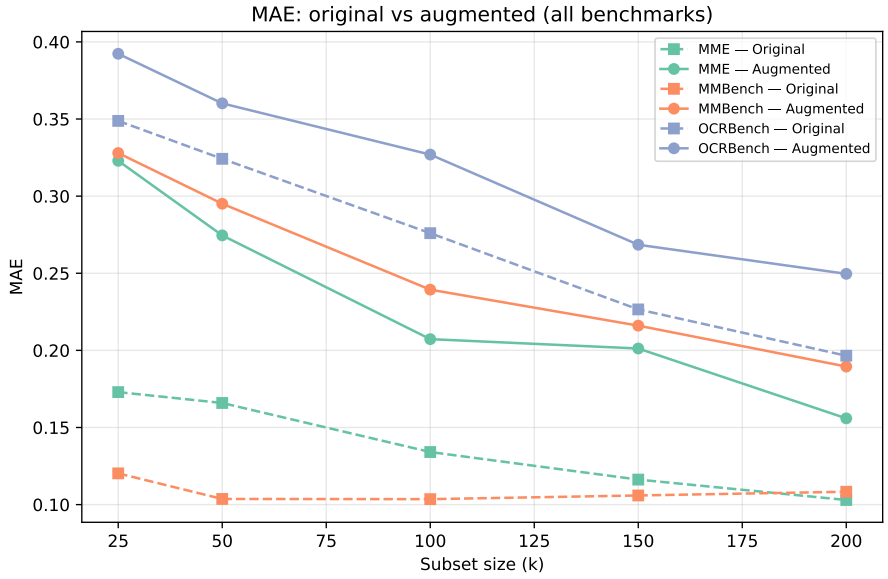


Figure 28: Score fidelity (MAE) as a function of subset size for original and formulation-augmented benchmarks across MME, MMBench, and OCRBench.

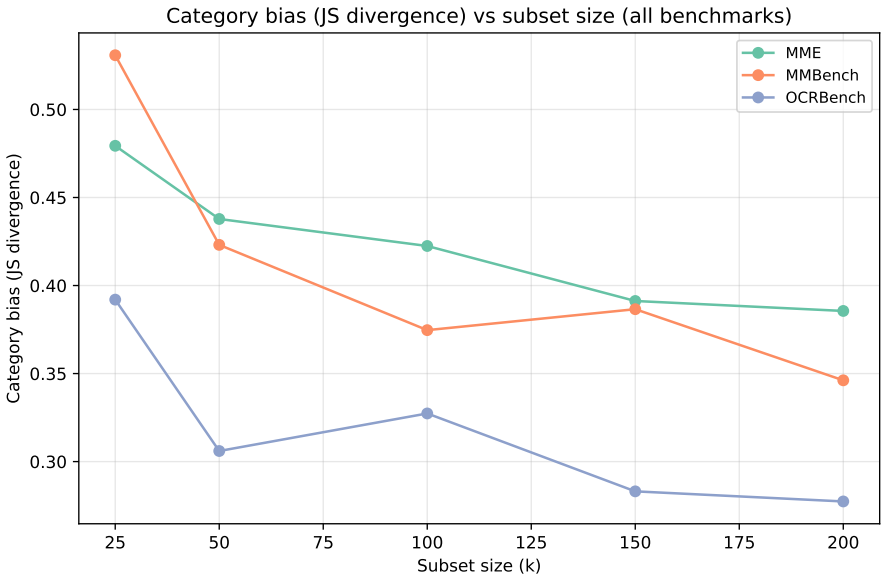


Figure 29: JS divergence between the category distribution of selected subsets and the full benchmark distribution as a function of subset size.

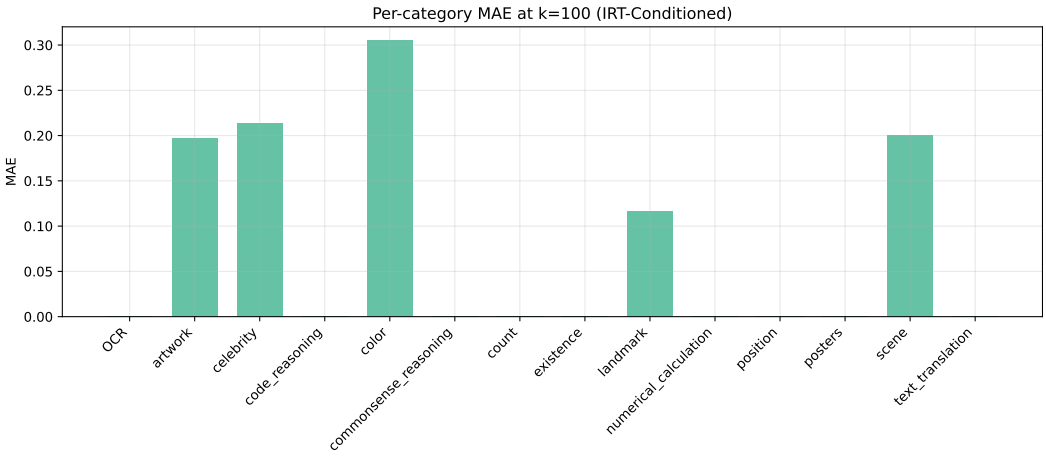


Figure 30: Per-category MAE at $k = 100$ for MME using formulation-aware IRT.

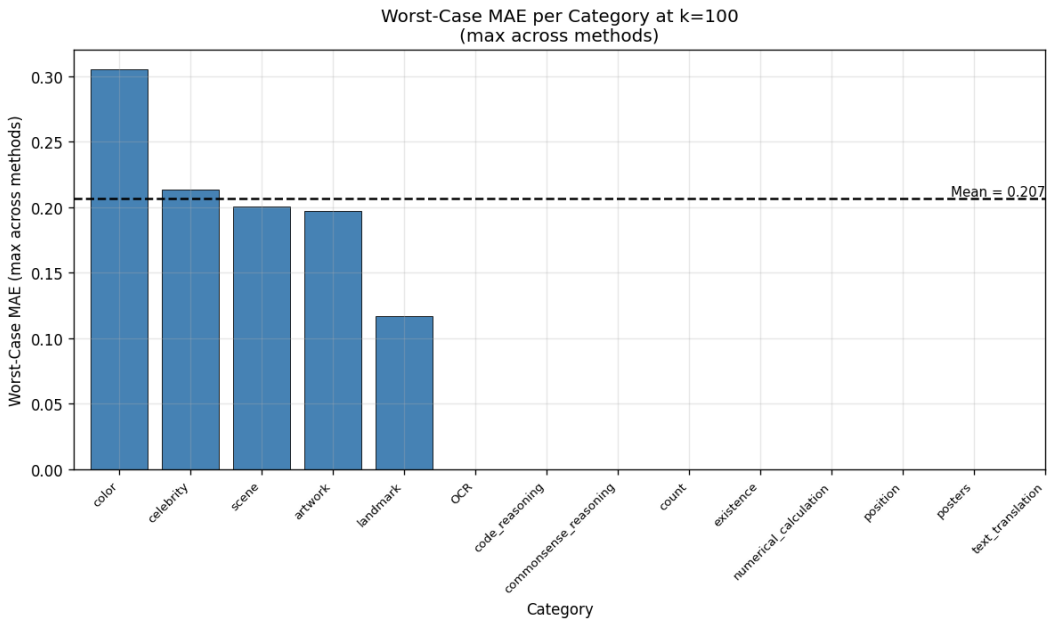


Figure 31: Worst-case category MAE at $k = 100$ for MME.

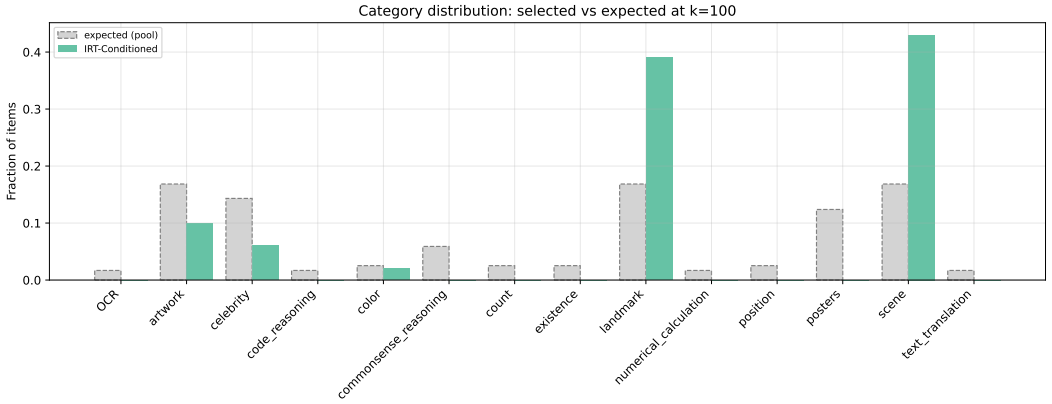


Figure 32: Category distribution of selected MME subset compared to the full benchmark distribution at $k = 100$.

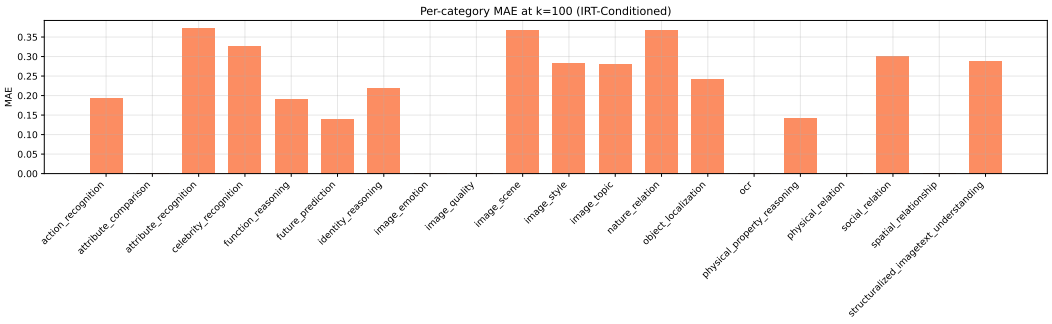


Figure 33: Per-category MAE at $k = 100$ for MMBench using formulation-aware IRT.

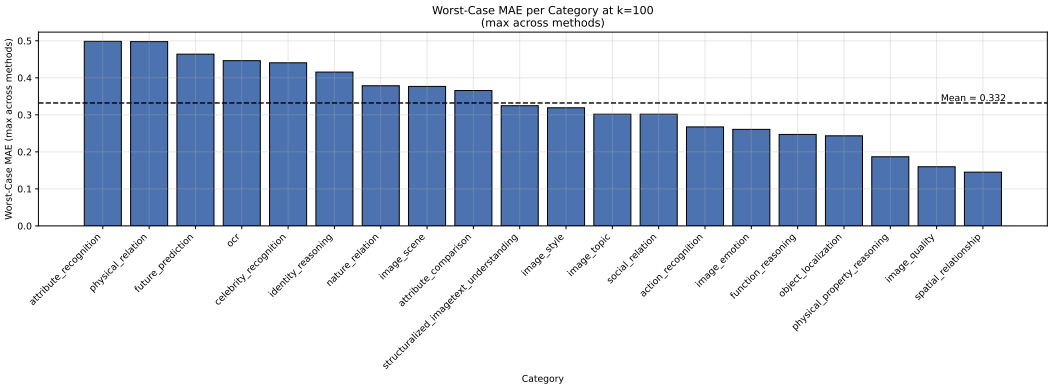


Figure 34: Worst-case category MAE at $k = 100$ for MMBench.

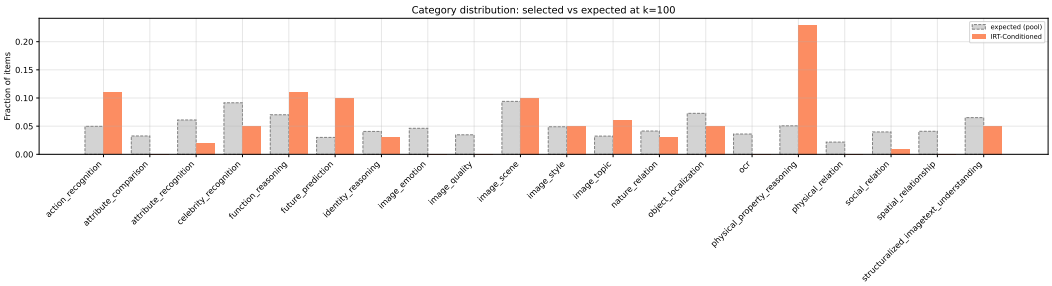


Figure 35: Category distribution of selected MMBench subset compared to the full benchmark distribution at $k = 100$.

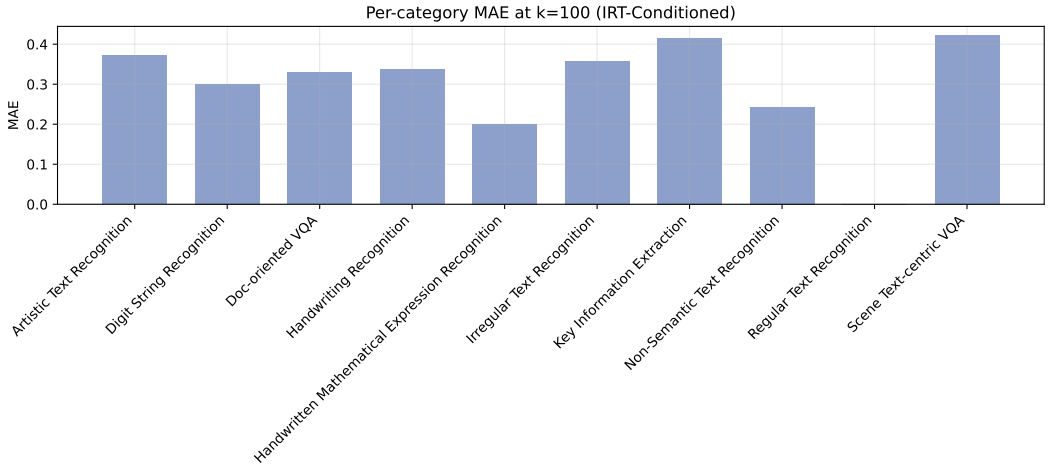


Figure 36: Per-category MAE at $k = 100$ for OCRBench using formulation-aware IRT.

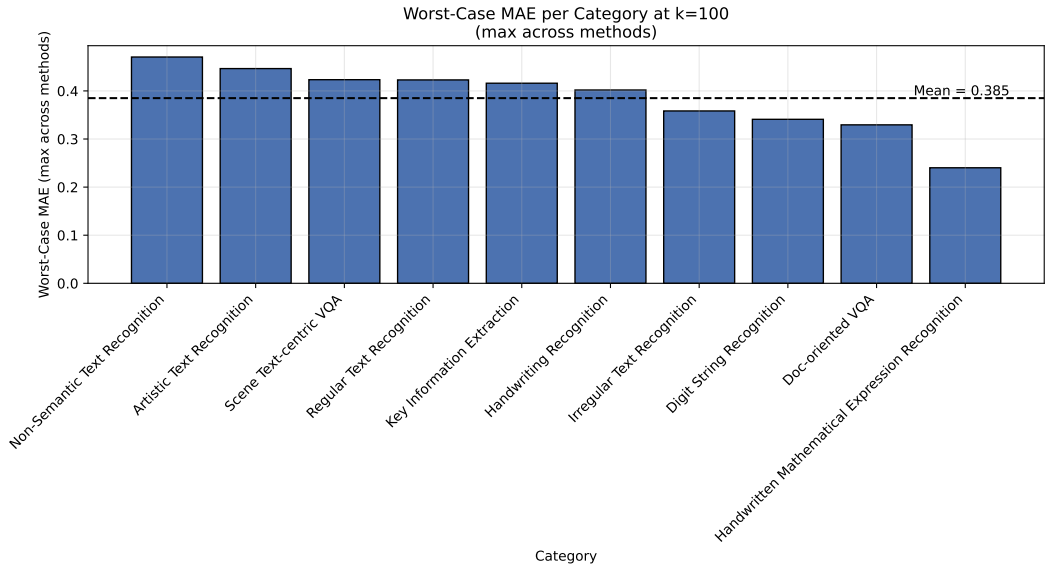


Figure 37: Worst-case category MAE at $k = 100$ for OCRBench.

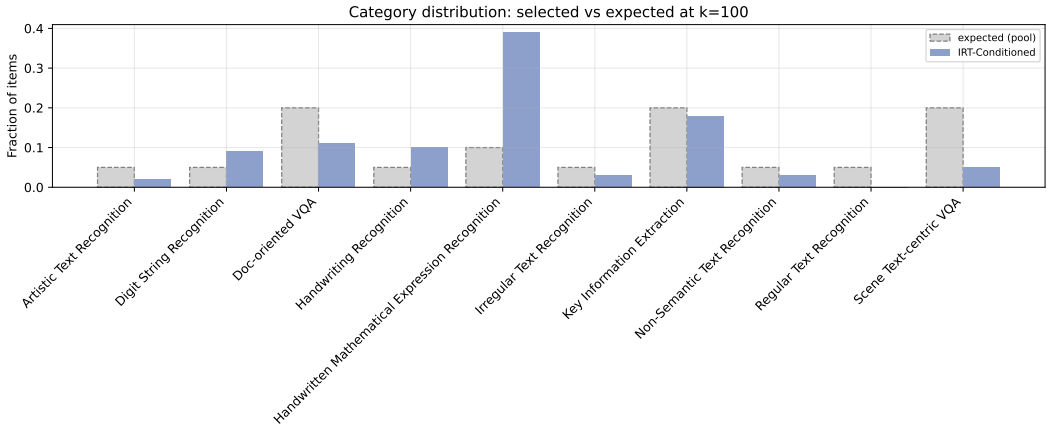


Figure 38: Category distribution of selected OCRBench subset compared to the full benchmark distribution at $k = 100$.