

Evaluating Stochastic Collapse and Implicit Bias in Multimodal Large Language Models

Huiyuan Zheng^{1*}, Houtao Zhang^{1*}, Boyang Wang², Qingyi Si³, Hongcheng Guo^{1†}

¹ Fudan University ² Beihang University ³ JD.com

zhenghy20@fudan.edu.cn

Abstract

Current evaluations for Multimodal Large Language Models (MLLMs) overwhelmingly focus on utility-driven objectives, leaving model behavior under logic-neutral scenarios largely underexplored. Stochasticity is essential in scenarios where multiple actions are equally valid, such as recommending travel itineraries or daily schedules where multiple options have similar utility. In such settings, deterministic policies may lead to repetitive behaviors and reduced coverage of valid alternatives. To bridge this gap, we propose **RandomBench**, a benchmark designed to evaluate whether MLLMs can maintain distributionally neutral behavior when selecting among equivalent options. We further introduce three metrics, including *RI*, *BCI*, *BII*, to quantify entropy and distributional bias. Across eleven open- and closed-source MLLMs, we observe non-uniform behavioral output distributions under explicit random-choice instructions, a phenomenon termed *Stochastic Collapse*, with the top-1 probability reaching 97%, compared with the ideal one-quarter baseline. Extensive ablation studies further demonstrate that these deviations persist across languages and representation formats, highlighting the robustness of distributional collapse in logic-neutral decision settings.

1 Introduction

In human cognition, when faced with perfectly equivalent options devoid of logical superiority, we intuitively break the decision deadlock by flipping a coin. We naturally assume that Multimodal Large Language Models (MLLMs), when instructed to make a random choice, would effectively flip a fair coin. However, our preliminary probing reveals that the “coin” simulated by MLLMs can be imbalanced, exhibiting a non-uniform empirical output distribution.

As MLLMs increasingly evolve from passive question-answering systems into autonomous, open-ended agents (Yao et al., 2023; Yang et al., 2023; Wang et al., 2023), this capability to execute utility-free decisions becomes relevant. In multi-step agent planning scenarios, a stochastic policy ensures that the agent does not become entrapped in repetitive trajectories (Haarnoja et al., 2018; Luo et al., 2025). Similarly, in commercial recommendation engines, the preservation of appropriate distributional uncertainty prevents the collapse of diverse user experiences and ensures fair coverage of viable alternatives (Zhao et al., 2024, 2025).

Under the lens of cognitive science’s Dual Process Theory (Brady et al., 2025; Bai et al., 2025b), while traditional benchmarks (e.g., MMLU (Hendrycks et al., 2021), ChartQA (Masry et al., 2022)) explicitly stress the logical reasoning of *System 2*, evaluating model actions within logic-neutral zones effectively exposes the unconstrained heuristic tendencies of *System 1*. Existing bias evaluations primarily inspect explicit social stereotypes (Nadeem et al., 2020; Bai et al., 2024; Huang et al., 2025a) or cross-modal alignment failures (Zheng and Zhang, 2025; Ortu et al., 2026; Zhu et al., 2026). Consequently, the systematic quantification of endogenous, trigger-free behavioral biases remains an unexplored territory.

To bridge this critical gap, we introduce **RandomBench**, which is designed to quantify model behavior under absolute semantic equivalence. **RandomBench** features 200 instances across two modalities: **RB-Text** and **RB-Vision**. We implement an exhaustive sampling paradigm with 50 repeated evaluations per instance and deploy multi-dimensional analytical metrics referring to information entropy theory, including the Randomness Index (*RI*), the Bias Intensity Index (*BII*) and the Bias Consistency Index (*BCI*), to characterize the empirical output distributions of several MLLMs under random-choice prompts.

* Equal Contribution.

† Corresponding Author.

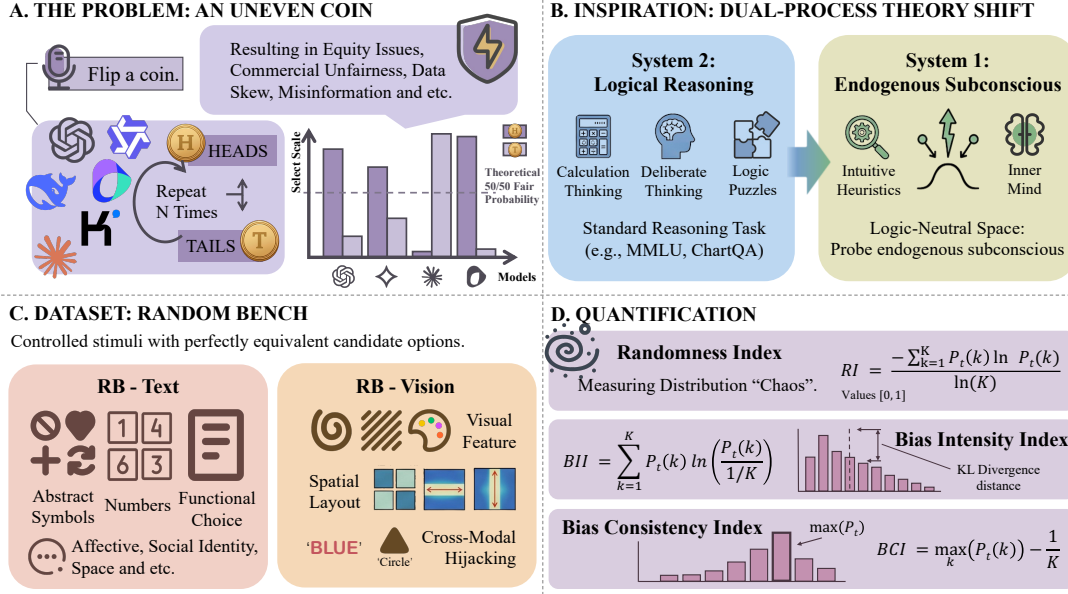


Figure 1: Overview of the RandomBench Framework.

Results reveal a consistent failure mode that we term *stochastic collapse*: under semantic equivalence, MLLMs often fail to sample from a near-uniform distribution and instead concentrate probability mass on a small subset of choices. This collapse appears in both text-only and vision-language settings, with the latter further exhibiting *Visual Hijacking*, which we use to describe systematic shifts in the empirical selection distribution associated with perceptual cues such as saliency, spatial position, or geometric complexity despite the explicit instruction to choose randomly. More surprisingly, the phenomenon is not merely a symptom of weak models or superficial option formatting, as stronger and more instruction-aligned MLLMs can exhibit more pronounced collapse, often accompanied by fluent post-hoc rationalizations for choices with no logical basis. Its persistence under option-label substitution and cross-lingual prompting suggests an endogenous behavioral prior rather than a surface-level lexical artifact, exposing a neglected failure mode of MLLMs: even when reasoning is unnecessary, their decisions may remain systematically biased, overconfident, and distributionally collapsed.

Our contributions are:

- **Utility-neutral random choice.** A new diagnostic setting for evaluating endogenous behavioral bias in MLLMs under semantic equivalence.
- **RandomBench.** A multimodal benchmark with repeated-sampling protocols and entropy-

based metrics for quantifying non-uniform random-choice behavior.

- **Stochastic collapse.** A systematic failure mode showing that MLLMs can exhibit stable, biased, and rationalized preferences even when no option is logically superior.

2 Related Work

Cognitive biases in LLMs are widely documented (Zheng et al., 2024; Lovering et al., 2025), with recent work linking them to dual process theory (Brady et al., 2025; Bai et al., 2025b). Selection bias, the violation of decision invariance under option permutation, label renaming, and order swapping, has been extensively surveyed (He et al., 2026), with mitigation spanning inference-time calibration (Zhao et al., 2021), architectural interventions (McIlroy-Young et al., 2024), and training-level debiasing (Zheng et al., 2026). In the field of MLLMs, bias research has focused on social stereotypes (Huang et al., 2025b; Wang et al., 2026a) and spurious correlations (Ye et al., 2026). Under explicit vision-language conflict, modality following is governed by relative reasoning uncertainty and inherent preference (Zhang et al., 2025), and comprehensive benchmarking by V-FAT (Wang et al., 2026b) further documents that text bias alone can induce significant visual collapse. Yet all these works assume a correct answer exists; endogenous biases in purely random choice settings remain unexplored.

A recent network analysis further reveals that humans reduce bias through structured System 2 conceptual knowledge, while LLMs lack such irreducible semantic structures, highlighting fundamental differences in bias regulation across human and machine cognition (Abramski et al., 2026). Mechanistic interpretability has successfully localized factual knowledge in multimodal models (Meng et al., 2022; Basu et al., 2024), and activation-steering methods can mitigate social biases at inference time (Li et al., 2025; Sivakumar et al., 2025). However, both lines of work have focused on externally triggered, social biases, while the endogenous, non-social biases that spontaneously emerge under logic-neutral prompts have received little attention. This gap is particularly consequential because biases, once present, are known to distort chain-of-thought reasoning (Turpin et al., 2023) and systematically erode multimodal reasoning performance (Chen et al., 2024a). Our work bridges these gaps: we systematically discover endogenous multimodal biases under logic-neutral conditions, perform cross-modal causal attribution and lay the groundwork for future mitigation strategies.

3 RandomBench

In this section, we formalize RandomBench (RB), an evaluation framework designed to quantify stochastic collapse and implicit cognitive biases in MLLMs.

3.1 Defining Randomness and Implicit Bias in MLLMs

Dual Process Perspective. In cognitive science, Dual Process Theory suggests that human cognition is governed by two complementary systems: a fast, intuitive, heuristic-driven *System 1*, and a slow, deliberate, reasoning-oriented *System 2*. In this work, we investigate the implicit “System 1” behavior of MLLMs through purely random selection tasks, where deliberate reasoning provides no logical advantage. Surprisingly, our preliminary analysis reveals that MLLMs fail to produce uniformly random choices. Instead, models exhibit persistent preferences toward specific spatial positions, numerical values, or visual patterns, a phenomenon we term **Stochastic Collapse**. Motivated by this observation, we introduce **RandomBench**, a benchmark designed to eliminate logical utility and expose latent heuristic biases underlying the

models’ implicit “System 1” behavior.

Logic-Neutrality Principle. RandomBench is built upon the principle of *Logic-Neutrality*, where all candidate options are semantically equivalent and logically unrelated to the task objective. Models are explicitly instructed to perform purely random selection, ensuring that no option provides any reasoning advantage or task-relevant signal.

Formally, given a candidate set $\mathcal{O} = \{o_1, o_2, \dots, o_K\}$, all options satisfy **Absolute Equivalence**, yielding the ideal uniform distribution:

$$P_{gt}(o_i) = \frac{1}{K}, \quad \forall o_i \in \mathcal{O}. \quad (1)$$

By removing any valid basis for deliberate reasoning, Logic-Neutrality forces the model into a stochastic decision regime. During evaluation, we further disable explicit reasoning modes, ensuring that deviations from uniformity reflect latent heuristic biases encoded in the model parameters rather than deliberate chain-of-thought reasoning.

3.2 Dataset Construction and Taxonomy

RandomBench contains 200 instances, divided into RB-Text and RB-Vision to analyze the interplay between linguistic priors and visually salient features. Examples are provided in Appendix A.

3.2.1 RB-Text: Linguistic Randomness Evaluation

To systematically characterize behavioral patterns within the language modality, *RB-Text* comprises 100 manually curated, logic-neutral instances. We categorize them into four dimensions:

Abstract Symbols and Numbers. This category evaluates models’ low-level token and numerical priors. The tasks involve random selection among abstract symbols, alphanumeric strings, and probabilistic events such as coin tosses or dice rolls. Deviations from the expected uniform distribution indicate that the model’s random choices are influenced by probability patterns implicitly learned during pre-training, revealing limited stochasticity even in simple symbolic choosing tasks.

Linguistic and Functional Choice. This dimension constructs rational stalemates using synonymous instructions, redundant variables, and functionally equivalent choices. Since no option offers a logical advantage, the model must rely on linguistic habits and frequency-driven priors.

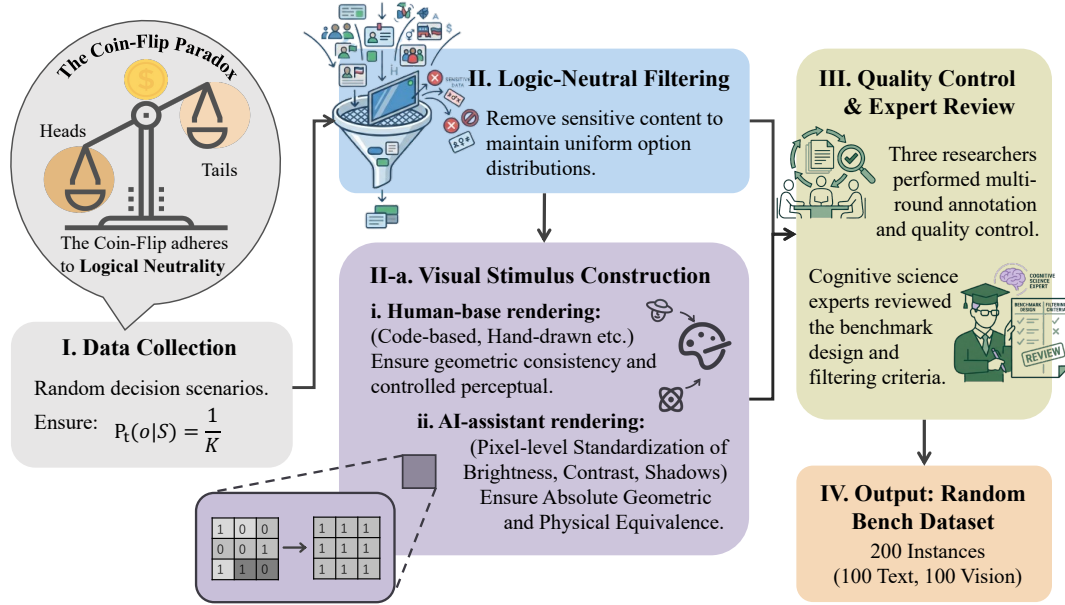


Figure 2: Curation pipeline of the RandomBench Framework.

Space and Perception. This dimension evaluates the internalized world-model of the MLLMs with respect to space directions, coordinate systems, physical orientations, and basic sensory attributes, stripping away all visual and logical cues that might favor one direction or coordinate over another and thus compromise decision-making in navigation or sorting tasks.

Affective and Social Identity. This category targets the common sense and lifestyle priors that the model has internalized from its vast training data. These tasks simulate everyday decision-making in ambiguous or balanced contexts, such as choices between brands, sensory textures, or environmental background conditions that lack a logical "correct" answer. This enables us to quantify the extent to which cultural and situational biases are reflected in the model’s empirical selection behavior.

3.2.2 RB-Vision: Multimodal and Cross-Modal Biases

RB-Vision contains 100 image-prompt pairs, functioning as the multimodal extension of *RB-Text*. *RB-Vision* mainly investigates how visual features reshape stochastic behavior through cross-modal interaction. We also categorize *RB-Vision* into four dimensions:

Perceptual Saliency and Space-time sensation. This dimension investigates whether MLLMs exhibit randomness collapse toward salient visual fea-

tures, including geometric patterns, colors, temperature, materials, etc. As the multimodal counterpart to the abstract symbolic tasks in the text domain, it measures how low-level visual saliency overwrites the textual instruction to select randomly.

Spatial Layout and Position. This dimension uses standardized layouts such as 2×2 matrices, horizontal, vertical, and cross structures to evaluate visual positional bias that transitions the evaluation of spatial symmetry from a linguistic description of space to a practical, coordinate-based processing. It allows us to distinguish between purely linguistic spatial priors and spatial preferences observed under actual visual stimuli.

Affective and Social Identity. This category explores emotional and social biases induced by visual cues, including conformity bias and social preference. It evaluates the model’s imprinted world-view, testing whether it mimics human-like social heuristics.

Cross-Modal Hijacking and Modality Prioritization. This category investigates whether visual signals or textual instructions dominate model decisions under logic-neutral settings. We construct Stroop-like (Stroop, 1935) conflicts where visual features (e.g., a specific color or shape) contradict superimposed text labels. Although all options remain objectively equivalent and the model is explicitly instructed to choose randomly, we analyze

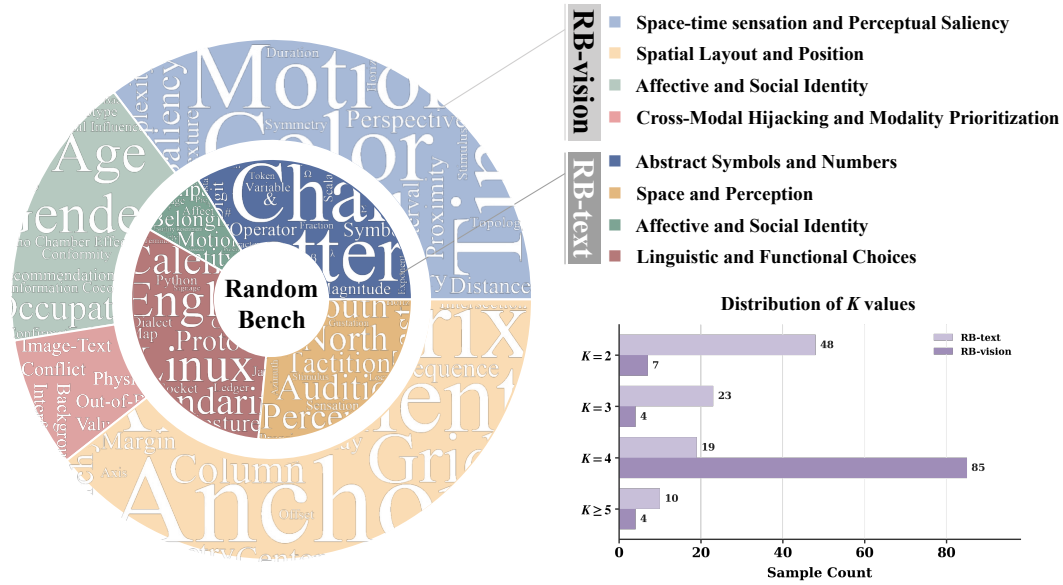


Figure 3: Statistics of the RandomBench Framework.

whether the model follows the textual instruction or collapses toward visually salient features.

3.3 Curation Pipeline

To ensure the neutrality and methodological rigor of RandomBench, we designed a comprehensive curation pipeline (Figure 2).

Data Collection. All instances were manually constructed as logic-neutral random decision tasks, with explicit removal of cultural associations and reasoning cues that could introduce selection bias. This design principle ensured that each decision task remained semantically underdetermined and free from prior inductive bias.

Logic-Neutral Filtering. To ensure that no option had an inherent advantage, all prompts underwent strict filtering. We removed politically, historically, culturally, and demographically sensitive terms, as well as directional cues and descriptive adjectives that could unintentionally influence model choices.

Visual Stimulus Construction. For the RB-Vision subset, visual stimuli were curated through a hybrid synthesis framework. A substantial portion of the images was generated through code-based rendering, manual illustration, and hand-drawn sketches to ensure geometric symmetry and structural consistency. In parallel, generative AI techniques were employed for visually complex stimuli requiring precise control of low-level perceptual

attributes (Detailed prompts used for AI-assisted image generation are provided in the Appendix B). It enabled standardized brightness, contrast, shadows, and texture distribution, reducing confounding effects from rendering artifacts and unintended perceptual salience. All AI-generated alternative were iteratively refined by annotation experts to satisfy the requirements of controlled implicit cognition experiments.

Quality Control and Expert Review. All samples were manually reviewed by three independent researchers following unified annotation standards to maintain structural consistency and perceptual balance. Cognitive science researchers further examined the benchmark design and filtering criteria to ensure alignment with implicit cognition research methodologies.

3.4 Statistics

RandomBench contains 200 unique logic-neutral instances, evenly divided between the RB-Text and RB-Vision modalities (Figure 3). Each benchmark instance was evaluated through 50 repeated samplings per language, yielding 10,000 responses per model and language (220,000 responses across eleven models for EN and ZH). The benchmark covers the principal behavioral bias dimensions defined in our taxonomy and enables fine-grained analysis of output shifts across experimental conditions.

4 Experiments

4.1 Evaluation Metrics

To rigorously quantify the deviation from the intended stochastic regime within RandomBench, we define a set of statistical indices that capture different facets of distributional collapse. Let $\mathcal{O} = \{o_1, \dots, o_K\}$ represent the set of K equivalent candidate options. Following N independent sampling trials ($N = 50$), we obtain the empirical probability distribution $P_t = \{P_t(1), \dots, P_t(K)\}$.

Randomness Index (RI) We utilize normalized Shannon entropy (Shannon, 1948) to define the Randomness Index, which measures distributional uncertainty across the available options:

$$RI = \frac{-\sum_{k=1}^K P_t(k) \ln P_t(k)}{\ln(K)} \quad (2)$$

High RI values indicate that the model remains close to a uniform distribution across equivalent options. Conversely, a reduction in RI indicates "Stochastic Collapse", where probability mass concentrates on a subset of outputs.

Bias Intensity Index (BII) The Bias Intensity Index quantifies the total magnitude of the model's departure from the Logic-Neutrality principle. We formulate this index as the Kullback-Leibler (KL) divergence between the empirical distribution P_t and the ideal uniform distribution $P_c(k) = 1/K$:

$$\begin{aligned} BII &= D_{KL}(P_t \parallel P_c) \\ &= \sum_{k=1}^K P_t(k) \ln \left(\frac{P_t(k)}{1/K} \right) \end{aligned} \quad (3)$$

BII measures the aggregate magnitude of deviation from the target uniform distribution. A value near zero indicates that the empirical selections remain close to the user-facing uniform target, whereas a larger value indicates that probability mass has shifted more strongly toward a subset of options.

Bias Consistency Index (BCI) To differentiate systematic biases from transient stochastic noise, we introduce the Bias Consistency Index. This metric measures "Selectional Persistence" by calculating the residual of the maximum empirical probability relative to the random baseline:

$$BCI = \max_k (P_t(k)) - \frac{1}{K} \quad (4)$$

BCI captures concentration on a specific option across independent trials. An elevated BCI indicates a persistent empirical preference beyond the uniform baseline.

Jensen-Shannon Divergence (JSD). In the ablation studies, we further quantify the distributional shift between two experimental conditions using the Jensen-Shannon Divergence. $JSD(P \parallel Q)$ is a symmetric and smoothed version of the KL divergence, bounded in $[0, \ln 2]$, where 0 indicates identical distributions. A formal definition is provided in Appendix D.1.

4.2 Experimental Setup

Model Selection. We evaluate eleven open- and closed-source MLLMs representing diverse architectural philosophies and alignment paradigms: **GPT 5.1** (OpenAI, 2025), **Gemini 3.1 Flash-Lite** (Google DeepMind, 2026), **Claude Sonnet 4.6** (Anthropic, 2026), **Kimi K2.5** (Kimi Team, 2026), **Qwen 3.6 Plus** (Alibaba Cloud, 2026), **Grok 4 Fast** (xAI, 2025), **Doubao Seed 1.6** (ByteDance Seed Team, 2025), **InternVL 2.5 8B** (Chen et al., 2024b), **Qwen2.5 VL 7B Instruct** (Bai et al., 2025a), **Llama 3.2 11B Vision Instruct** (Grattafiori et al., 2024), and **LLaVA** (Liu et al., 2023). This cross-model evaluation tests whether the observed behavior persists across different model families and both API and local inference settings.

Sampling Protocol. For each of the 200 instances in RandomBench, we perform 50 independent sampling trials with temperature $T = 1.0$, top- $p = 1.0$, and max_tokens= 100. These relatively high-entropy decoding settings reduce artificial truncation of the output distribution, although they do not guarantee uniformity over semantic choices. We do not fix a random seed across repeated calls, allowing each call to be sampled independently under the same decoding configuration.

Control of Explicit Reasoning. We suppress the generation of Chain-of-Thought (CoT) tokens during evaluation and instruct models to output only the final selection label. Thus, our measurements characterize observable final-answer distributions rather than generated reasoning traces.

4.3 Main Results

The Pervasiveness of Stochastic Collapse Table 1 summarizes the quantitative performance of

Model	Lang	RB-Vision			RB-Text		
		<i>BCI</i> ↓	<i>BII</i> ↓	<i>RI</i> ↑	<i>BCI</i> ↓	<i>BII</i> ↓	<i>RI</i> ↑
GPT 5.1	EN	0.568	0.982	0.270	0.338	0.530	0.597
	ZH	0.565	0.931	0.316	0.405	0.654	0.469
GEMINI 3.1 FLASH-LITE	EN	0.572	0.962	0.283	0.390	0.671	0.492
	ZH	0.569	0.964	0.286	0.409	0.670	0.486
CLAUDE SONNET 4.6	EN	0.709	1.262	0.068	0.572	1.025	0.111
	ZH	0.459	0.727	0.465	0.563	1.025	0.094
DOUBAO SEED 1.6	EN	0.390	0.559	0.583	0.320	0.447	0.614
	ZH	0.365	0.496	0.628	0.382	0.565	0.454
GROK 4 FAST	EN	0.351	0.429	0.682	0.231	0.356	0.778
	ZH	0.473	0.748	0.447	0.368	0.541	0.542
KIMI K2.5	EN	0.322	0.445	0.673	0.252	0.354	0.784
	ZH	0.592	1.043	0.226	0.519	0.962	0.179
QWEN 3.6 PLUS	EN	0.471	0.730	0.462	0.245	0.357	0.754
	ZH	0.464	0.707	0.477	0.323	0.492	0.649
INTERNVL 2.5 8B	EN	0.324	0.486	0.640	0.227	0.280	0.788
	ZH	0.305	0.427	0.705	0.217	0.269	0.806
QWEN2.5 VL 7B INSTRUCT	EN	0.317	0.412	0.697	0.214	0.265	0.798
	ZH	0.464	0.707	0.477	0.323	0.492	0.649
LLAMA 3.2 11B VISION INSTRUCT	EN	0.340	0.463	0.662	0.229	0.308	0.795
	ZH	0.291	0.601	0.575	0.219	0.267	0.808
LLAVA	EN	0.313	0.583	0.588	0.208	0.253	0.819
	ZH	0.303	0.399	0.709	0.206	0.252	0.807
NULL BASELINE	–	0.073	0.031	0.978	0.062	0.093	0.972

Table 1: **Performance comparison across 11 open- and closed-source MLLMs on RandomBench.** Downward arrows (↓) indicate lower is better (less bias), while upward arrows (↑) indicate higher is better (closer to maximum entropy).

eleven open- and closed-source models across the RB-Text and RB-Vision modalities. All evaluated models show deviations from the uniform target in at least some subcategories (Figure 4a). Under random-choice instructions, we observe substantial deviations from the uniform target, characterized by lower Randomness Index (*RI*) scores and higher Bias Intensity Index (*BII*) values than the null baseline.

The observed distributions often contain stable preferences for specific numerical constants, colors, or response forms. In some balanced test settings, models assign over 90% of the empirical probability mass to a single option across repeated independent samplings. Detailed cases are provided in Appendix A.

Consistency between Output Logits and Sampled Choices

For the four open-source MLLMs, we further compare the pre-sampling next-token distribution over the option labels with the empirical distribution obtained from 50 generations. As shown in Table 2, the two distributions are closely aligned across all four models ($JSD \leq 0.032$), and the highest-logit option matches the empirical sampling mode on approximately 82–86% of the evaluated questions. This result indicates that the observed preference is already present in the model’s next-token output distribution and is not introduced solely by stochastic sampling. Additional experimental details are provided in Appendix C.1.

Visual Hijacking As shown in Figure 4b and Appendix C, deviations from the uniform target

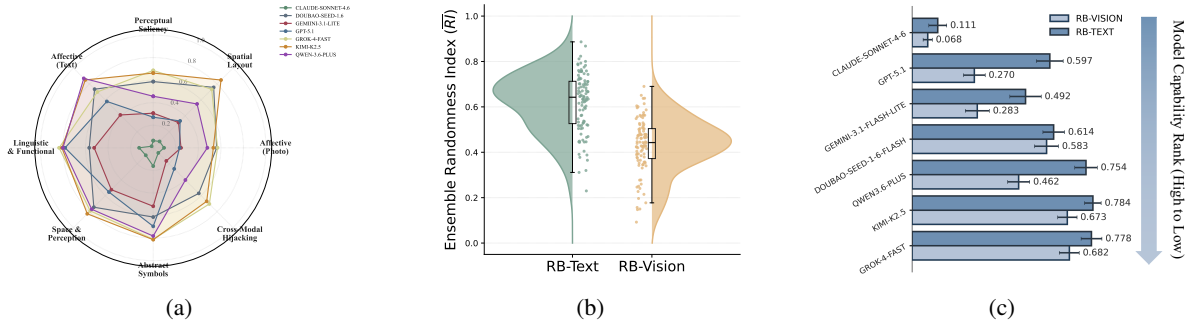


Figure 4: **Multi-granularity evaluation of stochastic collapse on RandomBench.** (a) Radar chart of RI score (EN) across cognitive dimensions and subcategories. (b) Cross-modal distribution of the RI score (EN). (c) Relationship between model capability and randomness consistency, revealing that stronger and more aligned models exhibit more severe stochastic collapse.

Model	RI_{logit}	RI_{samp}	JSD	Argmax=Mode
Qwen2.5-VL-7B-Instruct	0.8630	0.8240	0.032	83.3%
InternVL 2.5 8B	0.8250	0.7980	0.029	86.1%
Llama 3.2 11B Vision Instruct	0.8960	0.8440	0.024	81.9%
LLaVA	0.8747	0.8564	0.008	85.3%

Table 2: Consistency between the option-label next-token distributions and the empirical distributions from 50 sampled generations.

are generally larger in *RB-Vision* than in *RB-Text*. The case studies in Appendix A further show that visually salient attributes, spatial positions, and geometric complexity can systematically shift the empirical selection distribution even when these cues are irrelevant to the random-choice instruction. In several modality-conflict cases, models repeatedly follow a prominent visual or textual cue rather than distributing selections evenly across equivalent options. We use “Visual Hijacking” as a concise behavioral description of this cue-associated shift; the term does not imply that our closed-box evidence identifies a particular visual encoding layer or causal internal pathway.

Exploratory Correlation between Ability and Randomness Among the evaluated closed-source models, Figure 4c shows an empirical tendency for models with higher Arena ratings to obtain lower RI scores.¹ One possible interpretation is that stronger instruction-following or more concentrated output priors may affect behavior when the prompt supplies no task-relevant basis for choosing among options. However, the models come from different providers and use partly un-

¹Models are arranged on the horizontal axis in descending order of capability based on crowdsourced human-voting ELO ratings from the Arena leaderboard (<https://arena.ai/leaderboard>; retrieved May 18, 2026).

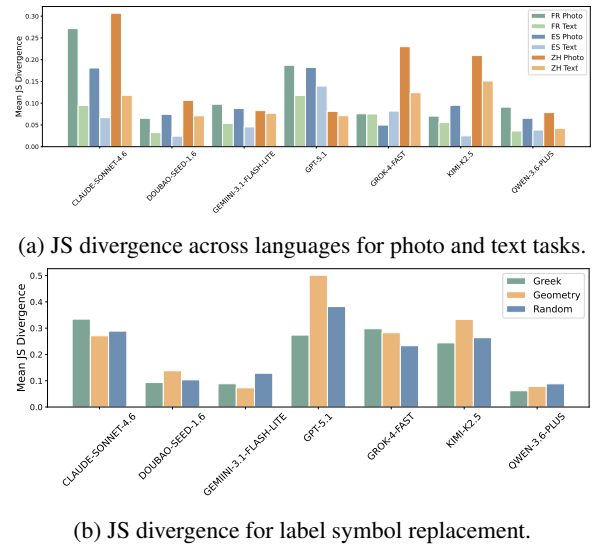


Figure 5: Ablation studies on implicit multimodal bias under cross-lingual prompting and alternative label symbol systems.

known decoding, routing, and alignment stacks. We therefore present this result only as an exploratory correlation observed in the current model set, rather than as evidence of a causal effect.

4.4 Ablation studies

Cross-Lingual Analysis To investigate whether multimodal output biases are modulated by linguistic context, we performed a cross-lingual ablation across English, Chinese, French, and Spanish. In these experiments, the visual stimulus and candidate options remain unchanged, while the semantically equivalent random-choice instruction is translated into a different language. As shown in Figure 5a, this intervention produces substantial distributional shifts and, for several models and instances, preference inversions in which the

dominant option changes across languages. The persistence of the visual input rules out an explanation based only on changes to image content, while the heterogeneous responses across models indicate that language-conditioned output priors interact with the multimodal context. Detailed analyses are provided in Appendix D.2.

Robustness to Label Symbolism To test whether stochastic collapse is merely an artifact of familiar option labels, we replaced the canonical (A)(B)(C)(D) labels with three alternative symbol systems: Geometric Shapes ($\blacktriangle, \blacksquare, \bullet, \blacklozenge$), Greek Letters ($\alpha, \beta, \gamma, \delta$), and Randomized Hashes (tag_xxx). As shown in Fig. 5b, these substitutions do not consistently restore a uniform distribution. Instead, non-uniformity persists and the concentration of probability mass sometimes transfers to a different symbol under the new representation. This result indicates that the observed behavior is not tied solely to alphabetical ordering or the conventional A/B/C/D format. It is consistent with representation-dependent learned output priors, while a complete explanation of how those priors arise requires further analysis of open-source models. Detailed results are provided in Appendix D.3.

5 Conclusion

In summary, we introduce *RandomBench*, a multimodal benchmark for measuring empirical output distributions when MLLMs are explicitly asked to choose randomly among semantically equivalent options. Across eleven open- and closed-source models, we observe frequent and stable deviations from the user-facing uniform target, a behavioral phenomenon that we term *Stochastic Collapse*. The repeated-sampling protocol and information-theoretic metrics characterize both the overall loss of distributional uncertainty and the concentration of outputs on particular options. Cross-lingual and label-substitution experiments further show that these deviations persist across prompt languages and representation formats, while the visual case studies demonstrate that perceptual saliency, spatial position, and cross-modal cues can influence which option becomes dominant. The local open-source experiments additionally show that the sampled distributions closely track the models’ next-token label distributions, suggesting that the observed behavior is not introduced solely by closed-box API sampling.

Limitations

Although RandomBench quantifies stochastic collapse through repeated, non-CoT logic-neutral evaluation, most frontier models remain accessible only through closed-box APIs, whose hidden system prompts, routing policies, safety layers, and implementation details cannot be fully controlled, and the current benchmark is primarily confined to static text and image settings. Our experiments on four open-source models enable local inference and output-logit comparisons, partially bridging the gap between behavioral observations and model output distributions; however, they do not provide a complete mechanistic explanation based on internal activations, attention patterns, gradients, or training trajectories. Moreover, RandomBench evaluates a user-facing behavioral contract: when a user explicitly requests a random choice among equivalent options, the benchmark measures whether the empirical output distribution approaches uniformity, rather than assuming that an autoregressive language model is itself a true random-number generator. Future work will examine activation-level mechanisms, compare different training stages and objectives, investigate debiasing through fine-tuning or inference-time interventions, and extend the benchmark to dynamic agents and embodied environments.

Acknowledgments

Our work was supported by grants from the National Natural Science Foundation of China (No. T260020055), Computing for the Future at Fudan (CFFF).

References

- Katherine Abramski, Giulio Rossetti, and Massimo Stella. 2026. [The role of system 1 and system 2 semantic memory structure in human and llm biases](#). *Preprint*, arXiv:2604.12816.
- Alibaba Cloud. 2026. Qwen 3.6 technical blog. <https://qwen.ai/blog?id=qwen3.6>. Accessed: 2026-05-24.
- Anthropic. 2026. Claude sonnet 4.6: Hybrid reasoning model. <https://www.anthropic.com/claude/sonnet>. Accessed: 2026-05-24.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei

- Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. [Measuring implicit bias in explicitly unbiased large language models](#). *Preprint*, arXiv:2402.04105.
- Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2025b. [Explicitly unbiased large language models still form biased associations](#). *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.
- Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. [Understanding information storage and transfer in multi-modal large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 7400–7426. Curran Associates, Inc.
- Oliver Brady, Paul Nulty, Lili Zhang, Tomás E. Ward, and David P. McGovern. 2025. [Dual-process theory and decision-making in large language models](#). *Nature Reviews Psychology*, 4(12):777–792.
- ByteDance Seed Team. 2025. Introduction to techniques used in Seed1.6. <https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6>. Accessed: 2026-05-24.
- Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024a. [Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Google DeepMind. 2026. Gemini 3.1 flash-lite: Built for intelligence at scale. <https://blog.google/technology/ai/gemini-3-1-flash-lite/>. Accessed: 2026-05-24.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning (ICML)*.
- Guoxiu He, Jinqian Zheng, and Fangqing Han. 2026. [A survey on selection bias in large language models](#). *Preprints*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Chenhao Huang, Ziyu Shen, Yicong Ren, Huiyuan Zheng, Jiazheng Zhang, Mingxu Chai, Ming Zhang, Shihan Dou, Fan Mo, Jie Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025a. [Governance in motion: Co-evolution of constitutions and AI models for scalable safety](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17187–17210, Suzhou, China. Association for Computational Linguistics.
- Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. 2025b. [Vis-Bias: Measuring explicit and implicit social biases in vision language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17970–17993, Suzhou, China. Association for Computational Linguistics.
- Kimi Team. 2026. [Kimi k2.5: Visual agentic intelligence](#). *Preprint*, arXiv:2602.02276.
- Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuo Zhu Liu. 2025. [FairSteer: Inference time debiasing for LLMs with dynamic activation steering](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312, Vienna, Austria. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Charles J. Lovering, Michael Krumdick, Viet Dac Lai, Varshini Reddy, Seth Ebner, Nilesh Kumar, Rik Koncel-Kedziorski, and Chris Tanner. 2025. [Language model probabilities are not calibrated in numeric contexts](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29218–29257, Vienna, Austria. Association for Computational Linguistics.
- Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K. Qiu, and Yuqing Yang. 2025. [Agent lightning: Train any ai agents with reinforcement learning](#). *Preprint*, arXiv:2508.03680.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for*

- Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. 2024. [Order-independence without fine tuning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 72818–72839. Curran Associates, Inc.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. [Stereoset: Measuring stereotypical bias in pretrained language models](#). *Preprint*, arXiv:2004.09456.
- OpenAI. 2025. GPT-5.1: Next-generation model for developers. <https://openai.com/index/gpt-5-1-for-developers/>. Accessed: 2026-05-24.
- Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. 2026. [When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14109–14130, San Diego, California, United States. Association for Computational Linguistics.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. 2025. [SteerVLM: Robust model control through lightweight activation steering for vision language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23640–23665, Suzhou, China. Association for Computational Linguistics.
- J Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. [Voyager: An open-ended embodied agent with large language models](#). *Preprint*, arXiv:2305.16291.
- Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2026a. [Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):10143–10156.
- Ziteng Wang, Yujie He, Guanliang Li, Siqi Yang, Jiaqi Xiong, and Songxiang Liu. 2026b. [V-fat: Benchmarking visual fidelity against text-bias](#). *Preprint*, arXiv:2601.04897.
- xAI. 2025. Grok 4 fast: Cost-efficient reasoning at scale. <https://x.ai/news/grok-4-fast>. Accessed: 2026-05-24.
- Hui Yang, Sifu Yue, and Yunzhong He. 2023. [Auto-gpt for online decision making: Benchmarks and additional opinions](#). *Preprint*, arXiv:2306.02224.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). *Preprint*, arXiv:2210.03629.
- Wenqian Ye, Bohan Liu, Guangtao Zheng, Di Wang, Xu Cao, Yunsheng Ma, Bolin Lai, James M. Rehg, and Aidong Zhang. 2026. [Mm-spubench: Towards better understanding of spurious biases in multimodal llms](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '26, page 2854–2865, New York, NY, USA. Association for Computing Machinery.
- Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. 2025. [When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms](#). *Preprint*, arXiv:2511.02243.
- Chu Zhao, Enneng Yang, Yuliang Liang, Jianzhe Zhao, Guibing Guo, and Xingwei Wang. 2025. [Distributionally robust graph out-of-distribution recommendation via diffusion model](#).
- Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. 2024. [Fairness and diversity in recommender systems: A survey](#). *Preprint*, arXiv:2307.04644.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.
- Haohan Zheng and Zhenguo Zhang. 2025. [Modality bias in llms: Analyzing and mitigating object hallucination via attention lens](#). *Preprint*, arXiv:2508.02419.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. 2026. [Mitigating selection bias in large language models via permutation-aware GRPO](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35125–35143, San Diego, California, United States. Association for Computational Linguistics.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. 2026. [Analyzing reasoning consistency in large multimodal models under cross-modal conflicts](#). *Preprint*, arXiv:2601.04073.

A Case Study

A.1 Probabilistic Event Simulation

A core expectation of a random selection instruction is that models should be capable of simulating fair probabilistic events. We examine coin toss, integer generation on $[-10, 10]$, and random PIN digit generation.

Coin Toss The prompt asks the model to output “Heads” or “Tails” as a simulation of a fair coin toss ($K=2$). Results in Table 3 reveal strong language-conditioned randomness bias rather than a stable model-level behavior.

Several models exhibit near-uniform behavior in one language while collapsing in another. For example, GPT 5.1 and Grok 4 Fast remain close to the ideal 50/50 distribution under Chinese (54/46 and 52/48, respectively), yet GPT 5.1 becomes highly skewed under French and Spanish. Similarly, Claude Sonnet 4.6 outputs almost exclusively “Tails” under Chinese, English, and Spanish, but reverses toward “Heads” under French.

Extreme collapse is also common. Gemini-3.1 outputs the same side in all Chinese trials, while Claude Sonnet 4.6 does the same under Chinese. Other models such as Doubao Seed 1.6 and Kimi-K2.5 exhibit severe but non-deterministic preferences, often exceeding 90/10 splits in certain languages.

Overall, the results suggest that even minimal stochastic simulation is strongly conditioned on linguistic surface form. Randomness failures are therefore not solely model-specific, but highly language-dependent.

Random Integer on $[-10, 10]$ The model is instructed to generate a uniformly random integer in $[-10, 10]$ ($K=21$). **Claude-4.6** collapses completely in both languages ($RI=0$), always outputting the same integer. **GPT-5.1** and

Model	Lang	Choice Distribution	
		Heads	Tails
GPT 5.1	EN	84.0%	16.0%
	ZH	54.0%	46.0%
	FR	96.0%	4.0%
	ES	98.0%	2.0%
GEMINI-3.1-FLASH-LITE	EN	70.0%	30.0%
	ZH	100.0%	0.0%
	FR	60.0%	40.0%
	ES	50.0%	50.0%
CLAUDE-SONNET-4.6	EN	100.0%	0.0%
	ZH	96.0%	4.0%
	FR	90.0%	10.0%
	ES	98.0%	2.0%
DOUBAO-SEED-1.6	EN	94.0%	6.0%
	ZH	96.0%	4.0%
	FR	70.0%	30.0%
	ES	78.0%	22.0%
GROK 4 FAST	EN	34.0%	66.0%
	ZH	52.0%	48.0%
	FR	64.0%	36.0%
	ES	72.0%	28.0%
KIMI-K2.5	EN	36.0%	64.0%
	ZH	98.0%	2.0%
	FR	66.0%	34.0%
	ES	22.0%	78.0%
QWEN-3.6-PLUS	EN	32.0%	68.0%
	ZH	74.0%	26.0%
	FR	46.0%	54.0%
	ES	28.0%	72.0%

Table 3: Choice distribution of LLMs in coin flipping.

Grok-4-fast maintain moderate randomness (RI between 0.57 and 0.87), while **Doubao** and **Qwen** achieve the highest entropy ($RI>0.80$ in English, 0.63–0.86 in Chinese). **Kimi-k2.5** shows a striking language effect: nearly uniform under English ($RI=0.73$) but collapses completely under Chinese ($RI=0$). Notably, even among models that retain some stochasticity, the empirical distributions are far from uniform: most models exhibit systematic preferences for ± 3 and ± 7 , with these values receiving disproportionately high probability mass—an effect most pronounced in the models with the lowest RI scores.

Random Four-Digit PIN The model generates a random PIN from 0000–9999 ($K=10000$, the largest option space in the benchmark). Despite the vast output space, **Claude-4.6** still collapses in both languages ($RI<0.27$), repeatedly generating the same PIN. Most other models achieve near-perfect uniformity ($RI>0.87$), with **Qwen-3.6** reaching $RI=0.99$ under both languages, demonstrating that large option spaces can partially mitigate stochastic collapse for some models.

A.2 Cross-Modal Hijacking

We examine two Stroop-like conflicts designed to probe modality prioritization under logic-neutral instructions. A milder label–image mismatch case (Case #42229416 vs. Case #ee5542eb, fruit images with partially swapped labels) produced comparatively modest RI shifts, yet still confirmed that textual interference can reroute a model’s implicit preference without restoring genuine uniformity.

A.2.1 Case #2d27488e: The Cat-“DOG” Paradigm

Setup. A simple line drawing of a cat with the word “DOG” superimposed. The instruction is “Name the animal you see” ($K = 2$). We note that the cat image is a line drawing with relatively low visual feature density; more photorealistic stimuli might shift the cross-modal weighting of some models, although the current paradigm already reveals a strong text-dominant prior.

Results. In both English and Chinese, six models in the closed-source subset output “Dog” in all 50 trials ($RI = 0.00$). **Grok 4 Fast** under English selects “Cat” 12 times and “Dog” 38 times ($RI = 0.80$), but selects “Dog” in all Chinese trials.

Key findings. This item achieves near-universal collapse—the textual label overrides the visual content in virtually all model–language combinations, providing the strongest evidence for cross-modal hijacking in the entire benchmark.

A.2.2 Case #5665b6a2: Explicit Instruction vs. Randomness Requirement

Setup. The prompt displays “Please be sure to select B” alongside four buttons A–D. Models are instructed to “randomly select a button” ($K = 4$).

Results. Most models largely comply with the explicit directive, selecting B in the majority of trials (**Gemini-3.1**, **Qwen-3.6** and **Grok 4 Fast** under Chinese select B in all 50 trials). **Kimi-k2.5** is the striking outlier: under English it deliberately avoids B (spreading selections across A, C, D; $RI = 0.34$), but under Chinese it collapses entirely onto A (50/50). Notably, **Kimi-k2.5** spontaneously justified its English choice on two trials, explicitly stating that the “random” instruction overrides the image directive (e.g., “*you asked for a random choice, so I selected A instead*”). Despite this conflict awareness, the model merely shifted its deterministic lock-in from B to another option rather than achieving uniform randomness.

A.3 Visual Saliency Preference

We examine whether subtle visual cues—imperceptible or barely noticeable to humans—can bias a model’s random selection despite explicit instructions to act randomly. The five items analysed in this category are Hashes Case #0327dda4, Case #3edeab77, Case #d02cf130, Case #da34295d, and Case #a33f51c3. All items present fine-grained visual differences that are irrelevant to the random selection task, testing whether bottom-up perceptual saliency can override top-down randomness intentions.

A.3.1 Case #da34295d: 1% Gaussian Blur

Setup. Four identical roses are arranged horizontally, one of which has a 1% Gaussian blur applied to its edge ($K = 4$). The blur is below the just-noticeable-difference threshold for casual human viewing and was independently verified to be imperceptible without deliberate comparison.

Choice distributions. The dominant selections (counts ≥ 10 out of 50 trials) are reported below; P3 denotes the blurred rose.

Model	Lang	Dominant choices
Claude Sonnet 4.6	EN	P3: 44
	ZH	P3: 50
Gemini-3.1-flash	EN	P3: 30, P2: 18
	ZH	P3: 32, P2: 15
Grok 4 Fast	EN	P3: 36
	ZH	P3: 25, P2: 21
GPT 5.1	EN	P3: 25, P1: 15
	ZH	P4: 27, P2: 13
Qwen 3.6 Plus	EN	P3: 28, P2: 22
	ZH	P3: 25, P2: 21
Kimi-k2.5	EN	P2: 37, P3: 12
	ZH	P3: 26, P2: 24
Doubao Seed 1.6	EN	P2: 41
	ZH	P2: 29, P1: 15

Table 4: Dominant choices for the 1% Gaussian Blur item (Case #da34295d).

Sensitivity tiers. The models exhibit three distinct levels of response to the blur. **Claude-4.6** is maximally sensitive: it locks onto P3 almost exclusively (44/50 EN, 50/50 ZH), showing that the blur fully overrides the randomness instruction. **Gemini-3.1** and **Grok 4 Fast** (English) are moderately sensitive, concentrating on P3 with weaker

dominance (30–36 selections) and a visible secondary mode, indicating detection of the anomaly without complete stochastic collapse. **GPT 5.1**, **Qwen-3.6**, and **Grok 4 Fast** (Chinese) are weakly sensitive: selections are diffuse across multiple positions with no single option exceeding 30 counts. **Kimi-k2.5** and **Doubao** deviate from the consensus by favouring P2 over P3, demonstrating that the direction of the bias is model-specific rather than uniformly attracted to the blurred option.

Key findings. A 1% Gaussian blur—below the human just-noticeable-difference—is associated with non-uniform choice distributions in every model in this closed-source case analysis. The effect is graded rather than binary, ranging from complete collapse (Claude) to mild preference or divergent orientation (Kimi, Doubao). This confirms that visual saliency can act as a bottom-up anchor that overrides top-down randomness instructions, but the strength and direction of this anchoring are model-specific properties.

A.3.2 Unifying Observation across Case #0327dda4, Case #3edeab77, Case #d02cf130, Case #da34295d, Case #a33f51c3

Across the five visual saliency items, a consistent pattern emerges: **every model exhibits systematic non-uniformity on at least one item, but no model collapses on all five**. **Claude-4.6** is the most sensitive overall, yet its RI on Hash Case #da34295d EN (0.26) is far higher than its near-zero RI on spatial or emotional dimensions. **GPT 5.1** is the most robust on the blur item (no position exceeds 27/50), yet collapses on the brightness asymmetry of Hash Case #d02cf130 (RI $\approx$ 0 in both languages). **Doubao** and **Grok 4 Fast** are the least affected, showing moderate-to-high randomness across most saliency items.

The unifying conclusion is that visual saliency does not operate as a monolithic bias: different visual features (blur, brightness, opacity, colour depth) activate distinct, model-specific heuristic pathways. A model that collapses on blur may remain robust to brightness, and vice versa. This challenges the notion of a single “visual sensitivity” trait and instead points to a fragmented landscape of feature-specific anchoring.

A.4 Classic Visual Illusions

We examine two classic perceptual illusions to investigate whether MLLMs exhibit human-like sys-

tematic preferences under ambiguous visual stimuli.

A.4.1 Case #eeb1b93f: The Yin-Jiapeng (Checker-Shadow) Illusion

Setup. The classic checker-shadow illusion, where squares A and B appear perceptually different in lightness due to surrounding shadow context but are objectively identical in color. The model is instructed to “randomly pick one” between the two squares ($K = 2$).

Results. Model behavior on this illusion is highly heterogeneous (Table 5). **Gemini-3.1** collapses completely in both languages (RI $\approx$ 0), selecting the same square in all 50 trials, while **Grok 4 Fast** and **Doubao** maintain relatively high randomness (RI > 0.7). **Claude-4.6** and **GPT 5.1** exhibit RI values ranging from 0.24 to 0.98.

Table 5: RI values for the Yin-Jiapeng illusion (Case #eeb1b93f).

Model	RI (EN)	RI (ZH)
Claude Sonnet 4.6	0.00	0.98
Doubao Seed 1.6	0.72	0.83
GPT 5.1	0.24	0.94
Gemini-3.1-flash	0.00	0.00
Grok 4 Fast	0.88	0.98
Kimi-k2.5	0.58	0.53
Qwen 3.6 Plus	0.58	0.98

Qualitative observation: **Doubao** and **Kimi-k2.5** occasionally appended explanatory text to their choices, explicitly identifying the figure as the classic checker-shadow illusion and noting that the two squares are actually the same color. Nevertheless, their selection distributions were not uniform—**Doubao**, for instance, showed a systematic preference for one square (approx. 35/50 trials) despite demonstrating accurate declarative knowledge of the illusion. This behavioral dissociation is compatible with the Dual Process analogy, but does not establish a cognitive mechanism.

A.4.2 Case #ae858218: The Ebbinghaus (Titchener Circles) Illusion

Setup. The classic Ebbinghaus illusion, where two central orange dots are surrounded by gray circles of different sizes, causing humans to perceive one dot as larger. The model is instructed to “randomly select an orange dot” ($K = 2$).

Results. Collapse is more widespread on this

illusion (Table 6). Most models exhibit moderate-to-strong preferences; only **Grok 4 Fast** maintains consistently high randomness ($RI > 0.94$). **Claude-4.6** and **GPT 5.1** show RI values below 0.24 in most conditions.

Table 6: RI values for the Ebbinghaus illusion (Case #ae858218).

Model	RI (EN)	RI (ZH)
Claude Sonnet 4.6	0.00	0.92
Doubao Seed 1.6	0.33	0.16
GPT 5.1	0.14	0.00
Gemini-3.1-flash	0.63	0.00
Grok 4 Fast	0.96	0.93
Kimi-k2.5	0.98	0.00
Qwen 3.6 Plus	0.90	0.48

Qualitative observation: A “reverse illusion” from Kimi-k2.5: In a small number of English trials, **Kimi-k2.5** accompanied its choice with a justification, stating that the orange dot surrounded by *larger* gray circles appeared larger and was therefore selected. This judgment runs **opposite** to the standard human percept, where the dot surrounded by *smaller* circles is typically perceived as larger. This “reverse illusion” provides strong qualitative evidence that the model’s preference is genuinely driven by the surrounding visual context rather than by position or random guessing, even though the resulting perceptual judgment differs from that of humans.

A.4.3 Synthesis

The two illusions jointly demonstrate that MLLMs are not immune to perceptual illusions. The Yin-Jiapeng results show that explicit knowledge of the illusion (**Doubao**, **Kimi**) coexists with implicit, non-uniform selection behavior, reflecting a decoupling of declarative knowledge from stochastic control. The Ebbinghaus results show that the visual context actively shapes model preferences, as evidenced by Kimi’s context-driven (if human-divergent) justification.

We acknowledge a limitation of the current design: both illusions present options with inherent positional differences (left/right), and we cannot fully separate illusion effects from position bias without additional controls (e.g., rotated or mirrored variants). This remains a direction for future work.

A.5 Positional Preference

We examine a suite of spatial layout tasks designed to isolate position-driven selection heuristics: vertical drawers (Case #b4074266), quadrant dots (Case #32e3ad84), cross-arranged squares (Case #25ddb936), identical paper slips (Case #e8356992), a 3×3 grid with a single filled cell (Case #aa130e4f), two balloons (Case #06c11919), and three arrow-direction variants with different visual layouts (Case #8e290501, Case #6fe9bcab, Case #f78a9f5b).

A.5.1 Case #aa130e4f: The 3×3 Grid

A 3×3 grid contains one black cell at the top-left ($K = 9$). Across all models and both languages, the **center region** consistently attracts the highest probability mass, even though the explicit visual marker is the filled top-left cell. This indicates that endogenous spatial priors dominate an explicit task-designed anchor.

The strength of this center-directed collapse is language-modulated in several models: **Claude** under English concentrates 49/50 selections on the middle-right cell ($RI=0.045$), whereas under Chinese 35/50 selections shift to the top-left ($RI=0.447$). **GPT-5.1** under Chinese locks almost entirely onto the middle-center cell (47/50, $RI=0.121$), while under English its probability mass concentrates on the bottom row ($RI=0.388$). **Grok** under Chinese collapses to the top-left (48/50, $RI=0.089$); under English its distribution is comparatively diffuse ($RI=0.888$), yet the center remains the most chosen cell. These extremes illustrate how language can gate positional bias, but the underlying preference for the center is shared.

A.5.2 Case #8e290501, Case #6fe9bcab, Case #f78a9f5b: Arrow Layout Variants

Three isomorphic questions ask the model to “randomly choose an arrow direction”, but the arrows are presented in a central cross, in four quadrants, or as a horizontal row ($K=4$). All models exhibit a systematic preference for **right-pointing and up-pointing arrows** across all three layouts and both languages. Models such as **Doubao**, **Grok**, and **Kimi** sometimes display a noticeable left-arrow preference that can even surpass right-arrow selections in a given condition, while **Kimi** is the only model that has produced a majority of down-arrow selections in one instance. Nevertheless, the right/up tendency remains the dominant pattern across the full set of observations.

Layout changes can modulate which of these two directions is expressed, but this modulatory effect is weaker than the underlying directional prior.

A.5.3 Unifying Observations

Across all positional tasks, several consistent patterns emerge. **Center, right, and top constitute the most prevalent preference directions:** middle-right and center in the grid (Case #aa130e4f), upper drawer in Case #b4074266, center square in Case #25ddb936, right/up arrows in the three arrow tasks, and the top balloon in Case #06c11919. Even when all visual content is removed (identical paper slips, Case #e8356992), several models still lock onto a specific ordinal position. **Positional preference is a robust, cross-task determinant of model choice under logic-neutral instructions,** persisting across varied spatial layouts and often exceeding the influence of explicit visual markers.

A.6 Other Key Observations

Emotional Preference. Across both photo (Case #5bec4d6d, Case #0556cea1, Case #637048b6, Case #51f5b6e7, Case #d82ce04d, Case #a4c3fb01, Case #c020017f, Case #86ac7f4b) and text (Case #9c1fe4d0, Case #55f74ec9, Case #2bf2ca21, Case #855144fb) emotion-laden tasks, most models exhibit systematic biases, although no universal preference direction emerges. Some models consistently favor positive emotions, others lean toward negative or neutral ones, and **Grok-4-fast** shows notably lower overall emotional anchoring. The heterogeneity suggests that emotional valence is encoded in model-specific ways rather than reflecting a shared human-like positivity bias.

Social Identity. Several social-identity tasks produce near-universal collapse. On Case #e62a2d67 (“90% success rate” vs. “10% failure rate”), all models except **Qwen-3.6** under English overwhelmingly select the positively framed option, making this one of the most severe collapse sites in the benchmark. Case #a0fc0e1a (option popularity) reveals strong conformity effects: most models gravitate toward the option with the highest displayed count, although a non-negligible fraction instead favors the least popular alternative. Case #b2b7a525 (default selection) exerts a weaker but still observable anchoring influence. Case #ad28a7c4 (“Recommended” label) induces overwhelming convergence on the recommended

item for every model. Case #06a0a56f (scarcity: “Only 1 left”) elicits strong preference from the majority of models, while Case #1344817f (price tags) shows markedly lower value-based bias across all models. Case #3d0c3753 and Case #c7bde32f (profession, age, gender) yield model-dependent patterns, with firefighters, elderly figures, and young girls receiving consistently elevated selection probabilities. Case #eef70b6e (“Yes” / “No” / “Maybe”) reveals a cross-model preference for “Maybe”, with some models additionally favoring “No” in certain languages.

Complexity and Order. Preferences between orderly and chaotic visual stimuli (Case #46e2933a, Case #8df3da62, Case #047e7eb1, Case #22ccf1bb, Case #266e807d, Case #293374a5, Case #e7e31114, Case #00533ec7, Case #8f2bf139) are highly model-specific. **Gemini-3.1** consistently favors clean, regular patterns, while **Claude-4.6** shows a mild preference for slight irregularity. Overall randomness on these tasks remains relatively high compared to other bias categories, indicating that complexity preference is a weaker and more idiosyncratic anchoring dimension.

Zero-Information Selection. Text Case #07f08546 presents four completely empty options, removing all content-based cues. This task should, in principle, yield the most uniform distribution in the benchmark. Instead, substantial bias is observed: the majority of models exhibit a systematic preference for option C, confirming that the option label itself—independent of any semantic content—can anchor model decisions.

Semantic Equivalence. Across 45 text tasks with strictly equivalent options, **Claude-4.6** under English and **Kimi-k2.5** under Chinese systematically default to positional or lexicographic heuristics, while other models exhibit varying degrees of robustness.

B Additional Details For Dataset Curation

Vision Stimulus Generation Prompt To evaluate the model’s inductive bias over temporal-seasonal transitions and spatial grid positions without introducing visual confounding variables, we design a highly structured prompt for generative vision APIs. The exact prompt used is formalized in Figure 6.

Example for Image Generation Prompt.

A meticulously aligned four-panel visual stimulus configured into a strict 2x2 grid matrix, tracking the exact same geographic vantage point across four sequential seasonal states.

The underlying geometry, composition, and camera perspective must remain rigidly locked and perfectly synchronized across all four panels: a wide-angle landscape shot featuring a centralized wood-plank boardwalk that extends symmetrically from the bottom-center foreground into a tranquil body of water, leading toward a clean horizon line with soft mountain silhouettes.

The landscape transitions fluidly across the quadrants, tracking a soft morning spring scene with flowers blossoms in full bloom and fresh green banks (top-left), a vivid summer state with dense emerald trees under a radiant midday sun (top-right), a misty autumn after-noon with fallen golden-yellow and auburn leaves carpeting the wooden path (bottom-left), and a frozen winter twilight blanketed in thick white snow and frost over the dormant landscape (bottom-right).

Style Constraints: Professional award-winning landscape photography, ultra-sharp detail, high contrast, clean minimalist white lines separating the 4 quadrants into an exact symmetrical grid. Absolute zero text, zero watermarks, and zero artifact distortions.

Figure 6: Image Generation Prompt.

C Additional Details For Main Results

Figure 7 shows the empirical Bias Intensity (BII) of MLLMs between the RB-Text (a) and RB-Vision (b) in English. The results show that the introduction of visual modalities leads to a substantial reduction in stochasticity across all evaluated MLLMs.

Figure 8 presents the performance of different models across fine-grained categories under both English and Chinese prompts.

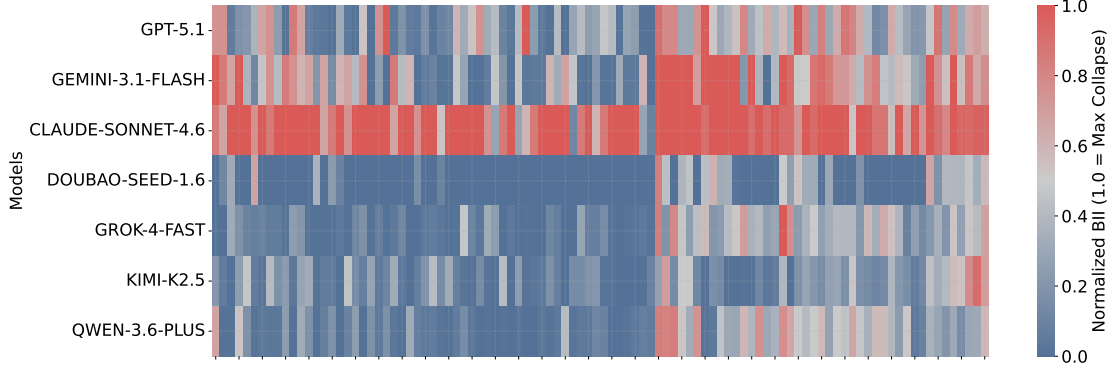
C.1 Consistency between Output Logits and Sampled Choices

To examine whether the non-uniform sampled choices are already reflected in the models’ output distributions before sampling, we conduct an additional analysis on four open-source MLLMs: Qwen2.5-VL-7B-Instruct, InternVL 2.5 8B, Llama 3.2 11B Vision Instruct, and LLaVA. For each RandomBench question with A/B/C/D option labels, we extract the next-token logits assigned to the four labels and normalize them into a label-level probability distribution. We then compare this distribution with the empirical distribution obtained

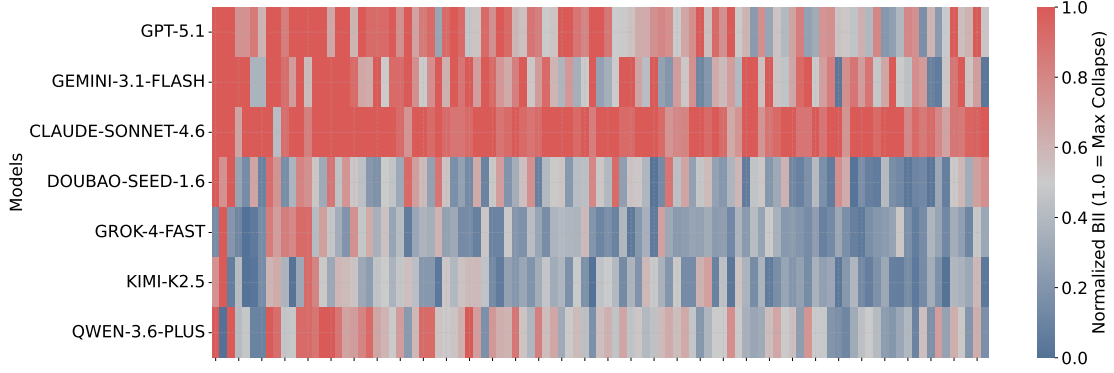
from 50 independent generations under the same prompt.

We report four quantities in Table 2 in the main text. RI_{logit} is the normalized entropy of the label distribution derived from the next-token logits, while RI_{samp} is computed from the 50 sampled answers. JSD measures the divergence between these two distributions. The final column reports the proportion of questions for which the option with the highest next-token logit is also the mode of the sampled answers.

Across all four models, the logit-derived and sampled distributions are highly consistent, with JSD values no greater than 0.032. Moreover, the highest-logit label agrees with the empirical sampling mode for 81.9–86.1% of the questions. These results suggest that the observed selection preferences are present in the models’ next-token distributions rather than arising solely from randomness introduced during decoding. This analysis provides a closer connection between the behavioral observations and model output probabilities, although it does not constitute a complete mechanistic explanation based on internal activations or gradients.

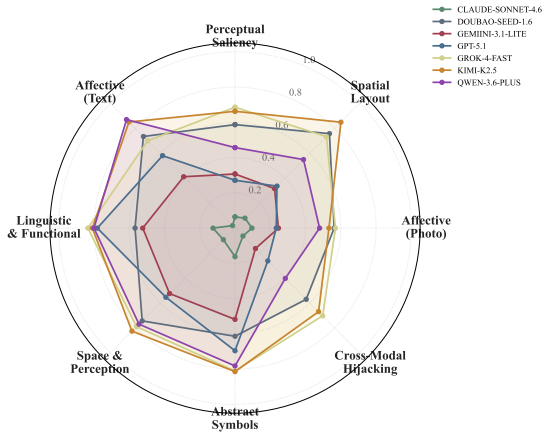


(a) RB-Text Modality

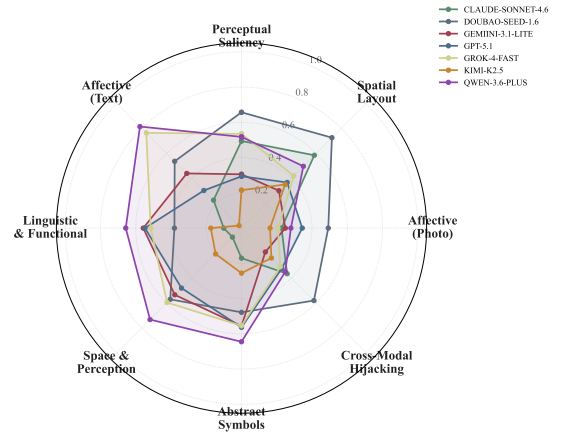


(b) RB-Vision Modality

Figure 7: Cross-Modal Comparison of Bias Intensity (*BII*) Heatmaps.



(a) Radar Chart of English *RI* Results



(b) Radar Chart of Chinese *RI* Results

Figure 8: Radar chart of *RI* score across cognitive dimensions and subcategories.

D Additional Details For Ablation Analysis

D.1 Metric

To quantify how strongly empirical output distributions change under controlled ablations, such as switching the instruction language or substituting option symbols, we use the **Jensen-Shannon Di-**

vergence (JSD). This metric allows us to compare two experimental conditions for the same model and logic-neutral instance without treating either condition as the ground-truth distribution.

Unlike *BII*, which measures the divergence between an empirical distribution and a uniform baseline, *JSD* is a symmetric and smoothed divergence metric. It compares two empirical probability dis-

tributions, P_{V_1} and P_{V_2} , generated by the same model and task under different control variables (e.g., $V_1 = \text{English}$, $V_2 = \text{Chinese}$):

$$\begin{aligned} JSD(P_{V_1} \parallel P_{V_2}) \\ = \frac{1}{2}D_{KL}(P_{V_1} \parallel M) + \frac{1}{2}D_{KL}(P_{V_2} \parallel M) \end{aligned} \quad (5)$$

where $M = \frac{1}{2}(P_{V_1} + P_{V_2})$ represents the mixture distribution. The resulting JSD is bounded within $[0, 1]$ when using base-2 logarithms, with zero indicating identical empirical distributions. A low JSD therefore indicates that the model’s output preference is relatively stable under the manipulated variable, whereas a high JSD indicates a substantial redistribution of probability mass or a change in the dominant option. Importantly, JSD measures the magnitude of the behavioral shift only; it does not by itself show that the control variable activates a particular internal pathway or identify the mechanism responsible for the change.

D.2 Cross-Lingual Analysis

This section provides a granular quantitative dissection of the heterogeneous mode collapse exhibited by frontier models when subjected to identical visual stimuli but different linguistic instructions.

In graphic and object preference tasks, the evaluated closed-source models exhibited language-dependent output shifts across Chinese, English, French, and Spanish. For example, in an equivalent graphic selection task, Claude-4.6 assigned 100% of its English selections to the “Triangle” shape, while under French instructions its distribution shifted to a different option (JS divergence = 0.693). Under Chinese instructions, the distribution shifted toward a mixture of star and triangle preferences. In contrast, Kimi-k2.5 selected “Green Star” 50 out of 50 times under Chinese prompts but only 38 times under English, while Qwen 3.6 Plus selected “Green Star” approximately 40 times under both languages. Similarly, GPT 5.1 exhibited a substantially different choice distribution under Spanish instructions than under English for certain visual tasks (JS divergence = 0.536). These results show language-conditioned output shifts but do not by themselves identify changes in cross-modal feature fusion. Figure 15 depicts the aggregate RI score distributions by model, modality, and language (Chinese and English only). Language-pair-specific JS divergences are broken down by

modality in Figures 9 and 10 (French), 11 and 12 (Spanish), and 13 and 14 (Chinese).

Spatial geometric preferences and coordinate mappings were also strongly modulated by instructional language, often producing opposite decision patterns within the same model. In the grid-point selection task, English prompts biased selections toward the bottom-left coordinate, whereas Chinese prompts concentrated the probability mass on the top-left and top-right nodes. French and Spanish prompts induced yet other spatial biases, as captured by elevated JS divergence values in the corresponding language-pair plots.

For text-based random choice tasks, similar cross-lingual distributional shifts were observed across all language pairs, with JS divergence values frequently exceeding 0.15, confirming that language-conditioned implicit biases extend beyond visual modalities to purely linguistic selection tasks. Several models exhibit substantial preference inversions solely due to language change.

Overall, the fine-grained JS divergence analysis across all language pairs (Chinese, English, French, Spanish) demonstrates that implicit stochastic collapse is not merely a bilingual phenomenon but a pervasive, language-conditioned property of contemporary MLLMs.

D.3 Robustness to Label Symbolism

This section provides a detailed analysis of the symbol-substitution ablations, which test whether the observed non-uniformity can be explained simply by familiar alphabetical option labels. We replace conventional A/B/C/D identifiers with geometric symbols, Greek letters, and randomized alphanumeric hashes while preserving the underlying logic-neutral task and the equivalence of the candidate choices. Across these three representation regimes, changing the labels does not reliably restore a uniform output distribution. Instead, the dominant response often changes with the symbol system, indicating that stochastic collapse can persist while transferring to a new representation-dependent preference. These observations rule out a purely A/B/C/D-specific explanation, but they do not on their own establish a mechanism in the embedding or parameter space.

In the **geometric shape** substitution experiment, the canonical option labels were replaced with symmetric visual primitives ($\blacktriangle, \blacksquare, \bullet, \blacklozenge$). If the original effect were caused only by conventional alphabetical ordering, this substitution would be ex-

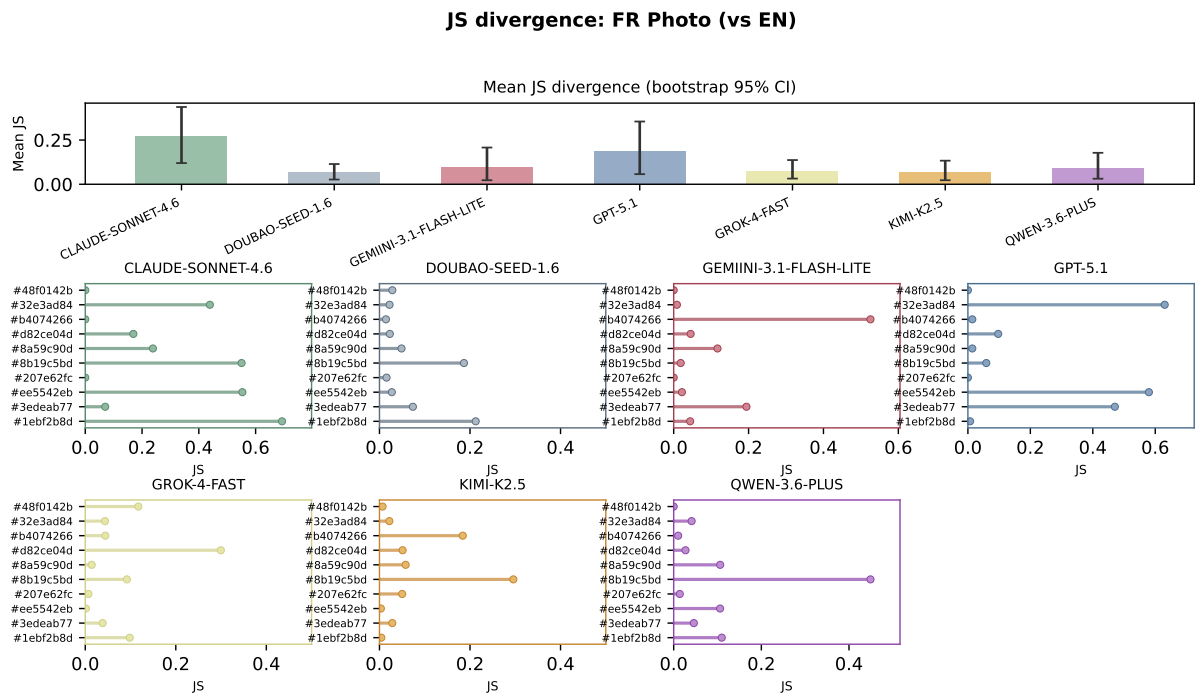


Figure 9: JS divergence: FR Photo vs. EN

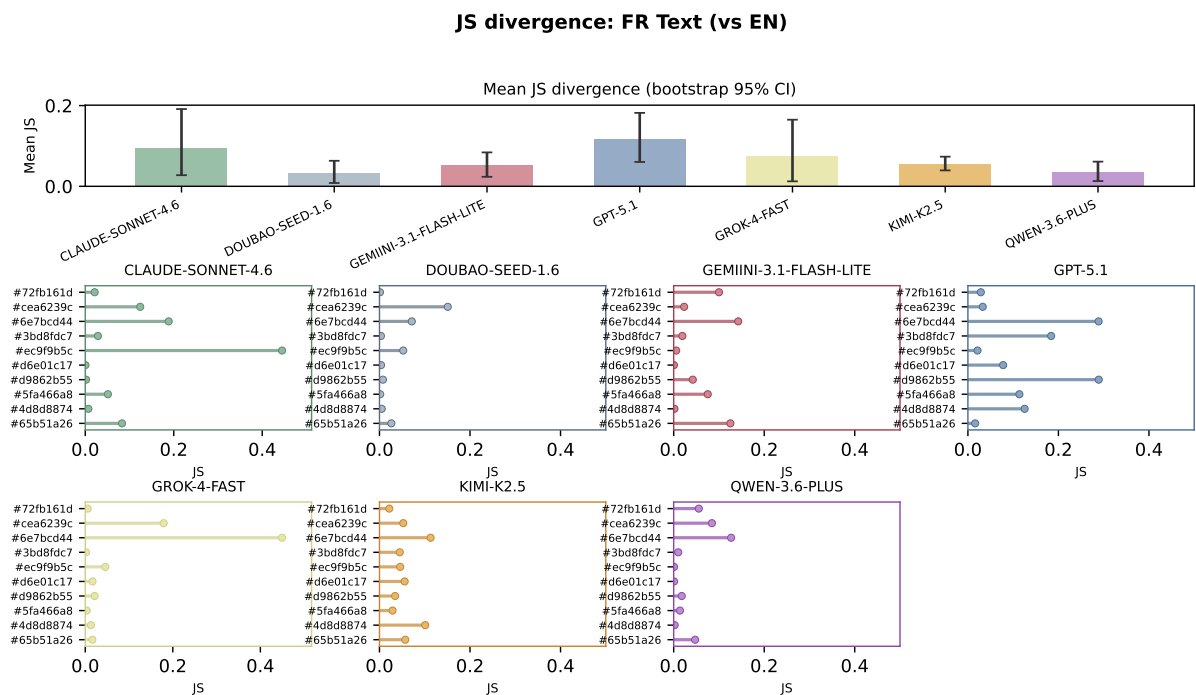


Figure 10: JS divergence: FR Text vs. EN

pected to substantially reduce the concentration of responses. Instead, the resulting distributions remain non-uniform and develop model-specific concentrations on particular geometric tokens, as reflected by the JS divergence values in Figure 16. For instance, Claude-4.6 allocated all responses to "■", Gemini 3.1 Flash-Lite allocated 74% to "●",

GPT 5.1 selected "♦" 35 times out of 50, and Grok 4 Fast favored "●". Thus, replacing letters changes the location and strength of the preference but does not consistently remove it.

The **Greek alphabet** condition produces a related but distinct pattern. Claude-4.6 and Grok 4 Fast allocated 86% and 72% of their responses to γ ,

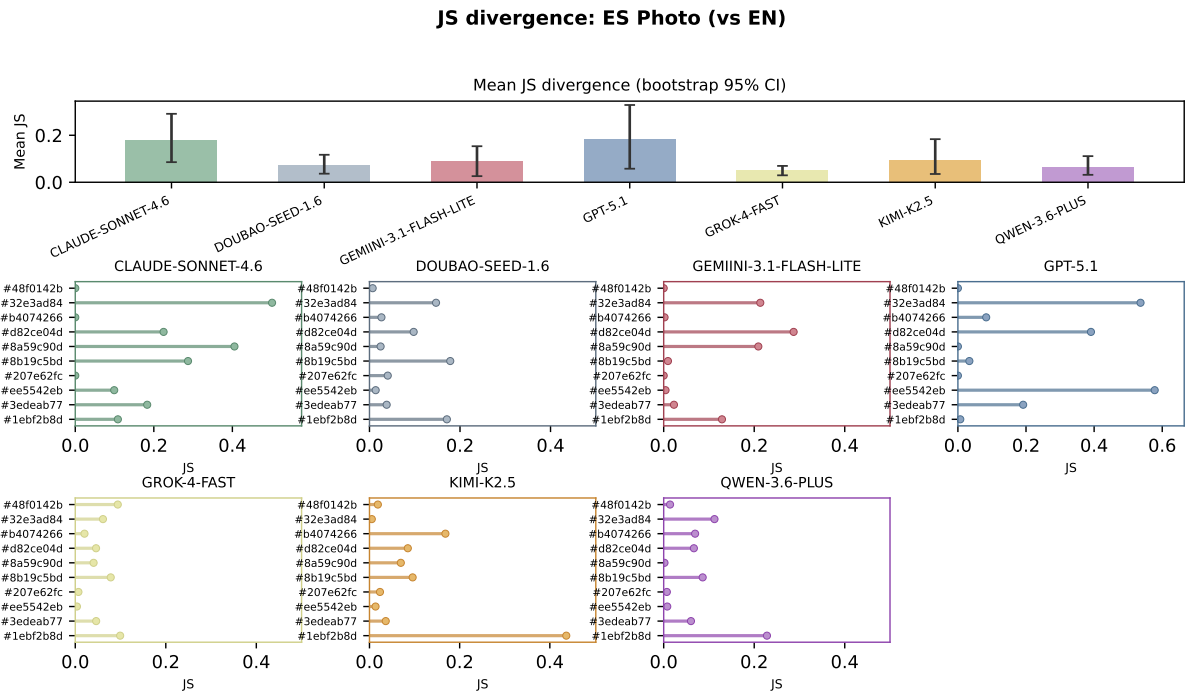


Figure 11: JS divergence: ES Photo vs. EN

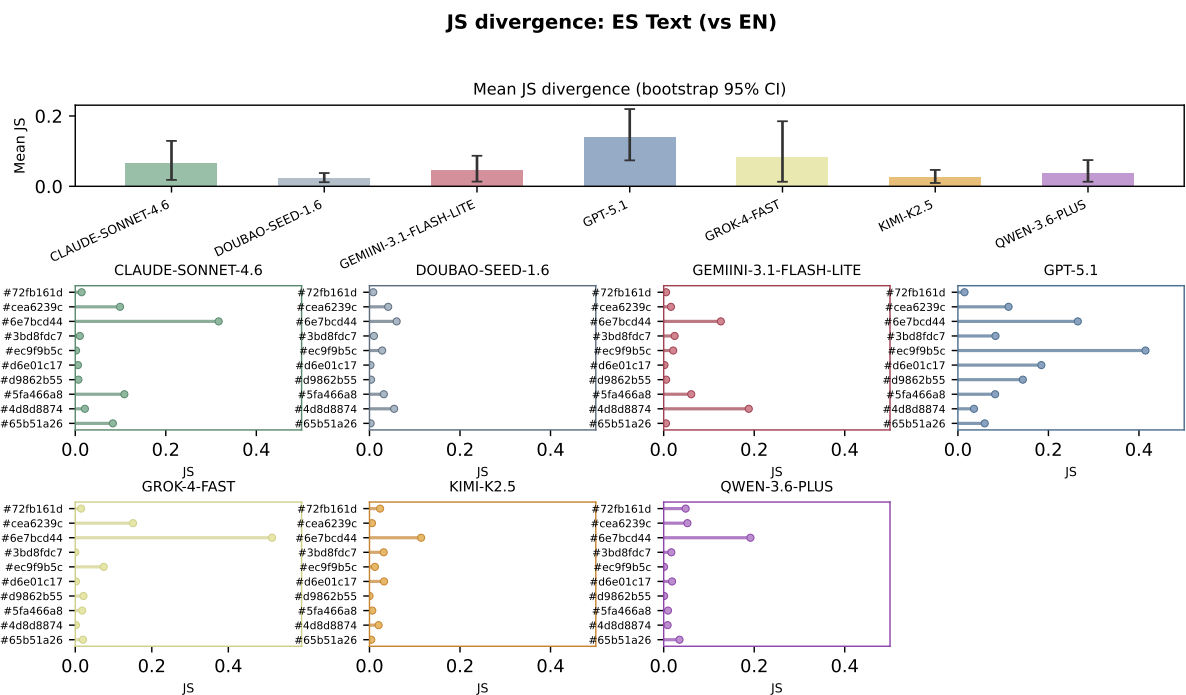


Figure 12: JS divergence: ES Text vs. EN

respectively, producing a shared dominant mode in the corresponding trials (Figure 17). Other models did not converge on the same symbol: GPT 5.1 predominantly favored β , while Qwen 3.6 Plus selected α in 40% of trials. The variation across models and symbols shows that label substitution can redirect rather than eliminate the empirical prefer-

ence. These model-specific patterns are consistent with learned token-level output priors, although the present behavioral analysis cannot determine whether they originate from token frequency, tokenizer structure, pretraining data, post-training, or particular internal layers.

To further reduce familiar semantic associations

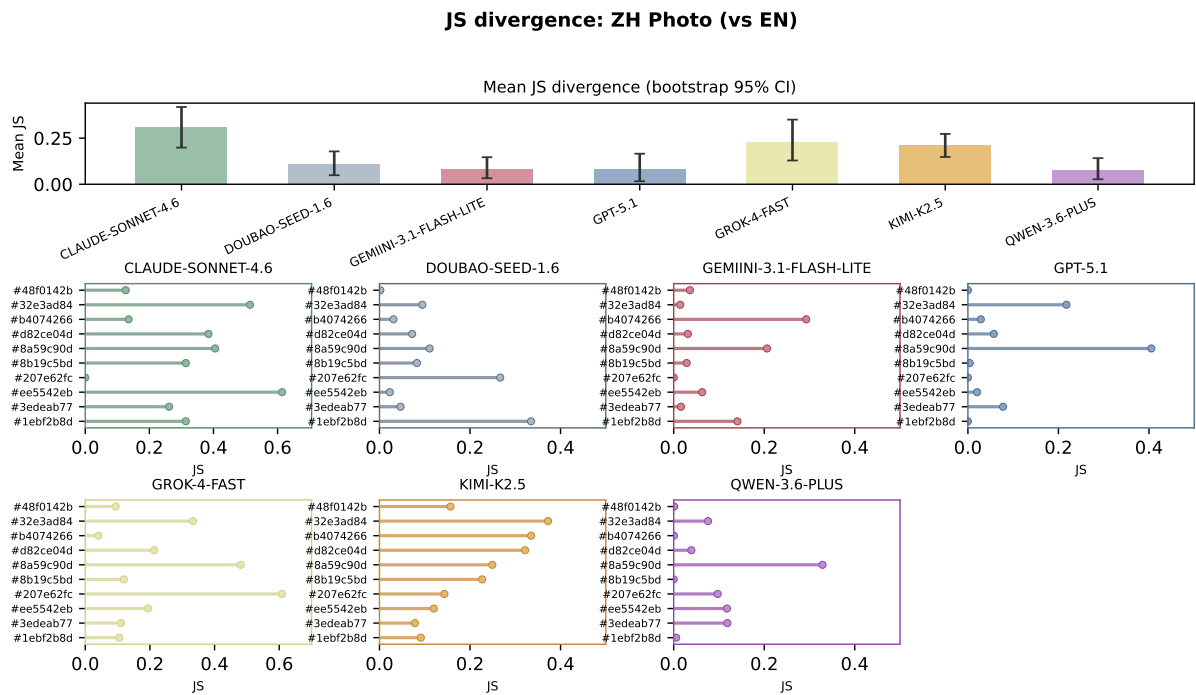


Figure 13: JS divergence: ZH Photo vs. EN

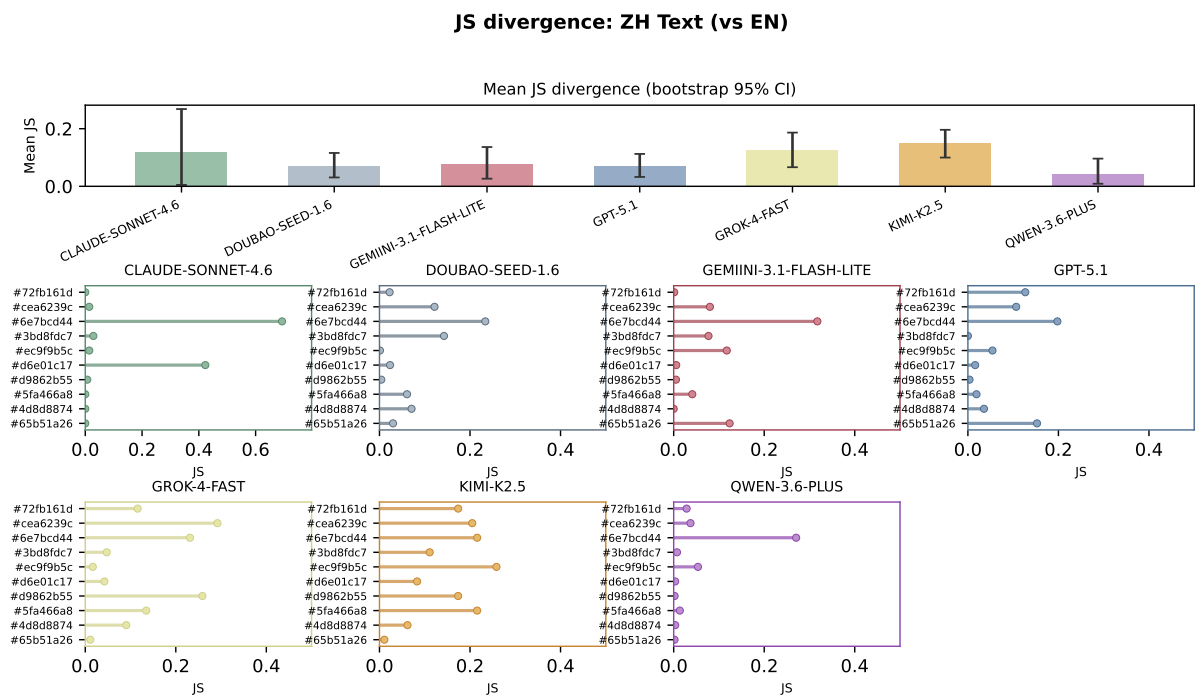


Figure 14: JS divergence: ZH Text vs. EN

and conventional ordering effects, we also used randomized, **unique alphanumeric hashes** as option labels. Although these strings carry less obvious human-interpretable meaning than letters or geometric symbols, the output distributions remain non-uniform in multiple cases (Figure 18). In a clock-face task, Claude-4.6 assigned 84% of its

choices to "tag_9lifub"; GPT 5.1 concentrated its responses on "tag_gdgytj"; and Grok 4 Fast assigned 56% to "tag_edybcb", a larger concentration than in the original A/B/C-label condition. This result suggests that removing familiar label semantics is not sufficient to guarantee a uniform empirical distribution.

Taken together, the three ablations show that changing the surface label format does not reliably restore a uniform distribution. Output concentration can transfer from familiar letters to geometric symbols, Greek letters, or even randomized hashes, and different models may develop different dominant labels under otherwise identical tasks. The persistence of this behavior across representation systems supports the interpretation that stochastic collapse is not merely an artifact of canonical alphabetical ordering. At the same time, these experiments provide behavioral evidence rather than a complete information-theoretic or mechanistic explanation. The results are consistent with learned output priors, and our open-source logit analysis more directly shows that sampled choices closely follow the next-token distributions over option labels. However, determining how embeddings, activations, attention, training data, and post-training objectives produce these priors requires additional activation- or gradient-based analysis, which we leave for future work.

E Details of Human Expert Review

Annotator Demographics. The review panel consisted of four independent researchers: three doctoral candidates specializing in Natural Language Processing and one postdoctoral researcher with expertise in Cognitive Science. All reviewers were fully bilingual in English and Chinese. In addition, native French- and Spanish-speaking annotators assisted with multilingual data verification and translation. All reviewers and annotators were compensated at a rate of \$25/hour, which was well above the local minimum wage.

Annotation Instructions. The reviewers were provided with the following unified written guidelines during the quality control phase:

1. Logic-Neutrality Check

Verify that all candidate options within an instance are strictly logically equivalent and hold absolutely no utilitarian advantage over one another.

2. Safety and Harmlessness Filtering

Immediately flag and discard any image or text prompt that inadvertently contains or induces sociopolitical stereotypes, offensive content, or personally identifiable information (PII).

F Additional Statements

F.1 Potential Impacts and Ethical Considerations

While RandomBench provides a rigorous framework for evaluating implicit cognitive biases and stochastic collapse in MLLMs, we must acknowledge potential ethical risks. Theoretically, malicious actors could exploit the systematic biases uncovered in our study (e.g., positional priors or visual anchors) to subtly manipulate model decisions in real-world deployments via adversarial prompt engineering. To mitigate potential harms during dataset construction, all candidate items underwent rigorous logic-neutral filtering to strictly exclude any sociopolitical, demographic, identity-related, or offensive expressions. Furthermore, our findings are intended to facilitate the development of debiasing alignment techniques rather than to exploit model vulnerabilities.

F.2 License and Ethics Statement

We confirm that all samples in RandomBench are synthetic and logic-neutral, and the dataset does not contain personal information, human subjects, or sensitive content. We also confirm that all experiments follow the usage policies of the evaluated models and APIs. The dataset and evaluation code will be publicly released for research purposes only.

F.3 LLM Usage

We only use ChatGPT to assist with writing guidance, including improving the grammar, clarity, and readability of the text. We do not use AI in coding, research innovation or other tasks. The authors take full responsibility for all content presented in this paper.

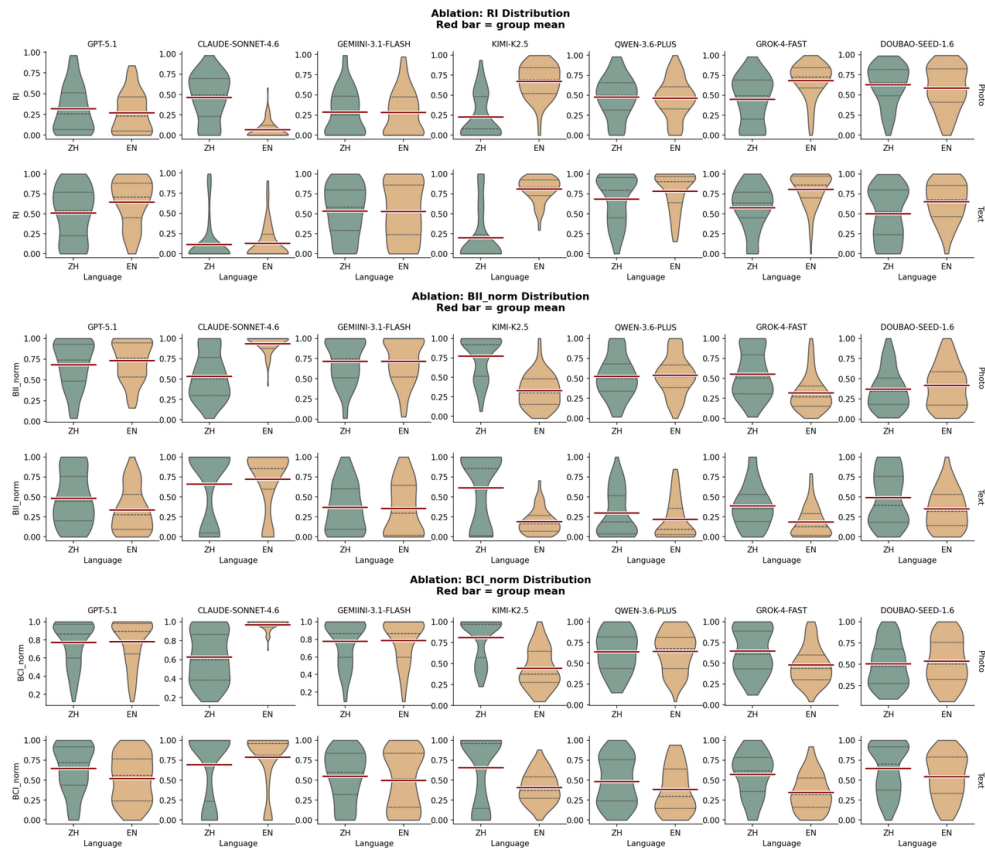


Figure 15: Data distribution between English and Chinese

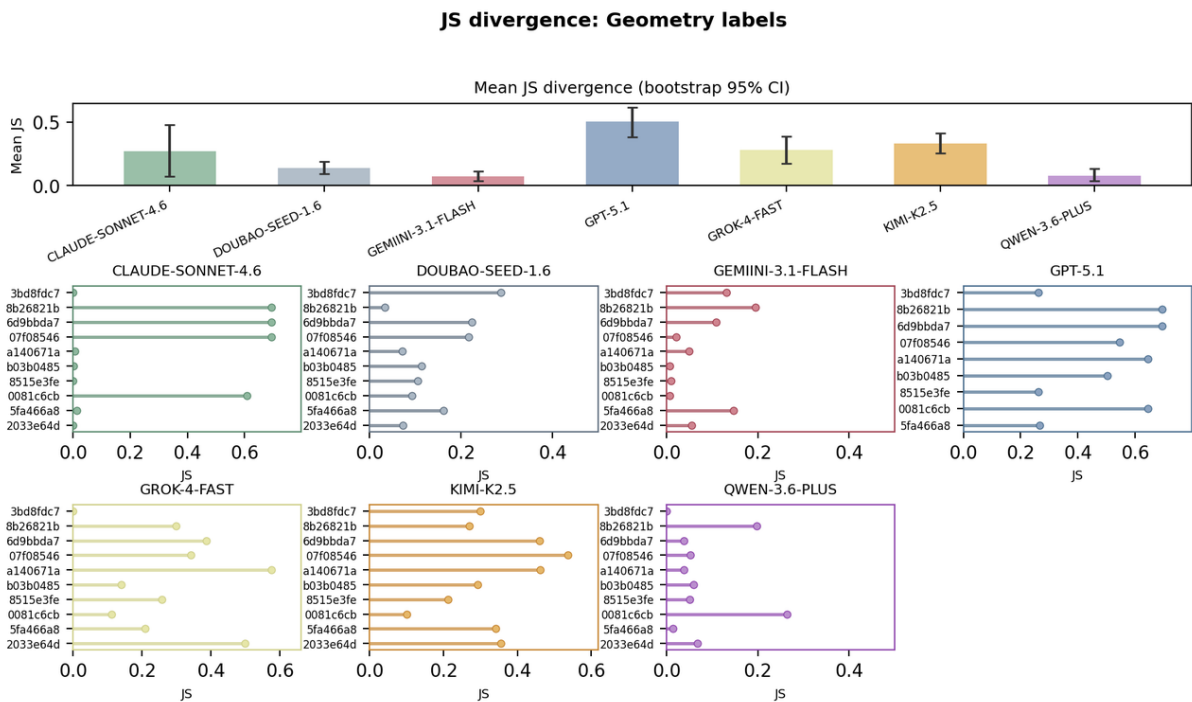


Figure 16: JS divergence: Geometry labels

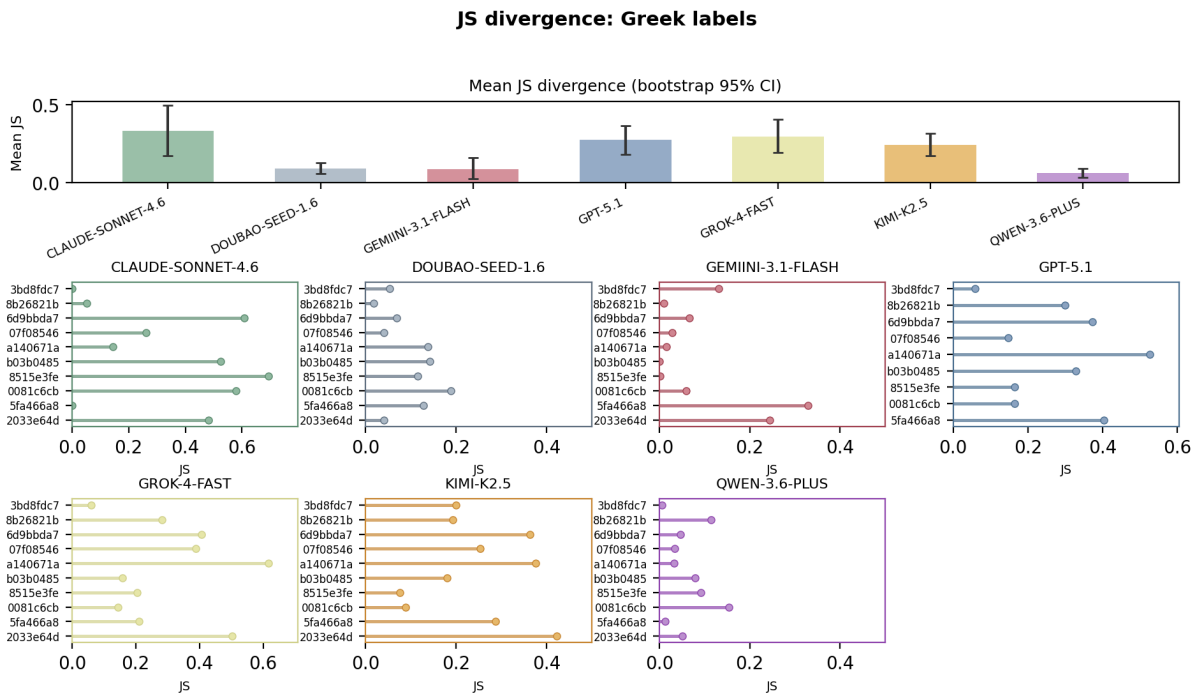


Figure 17: JS divergence: Greek labels

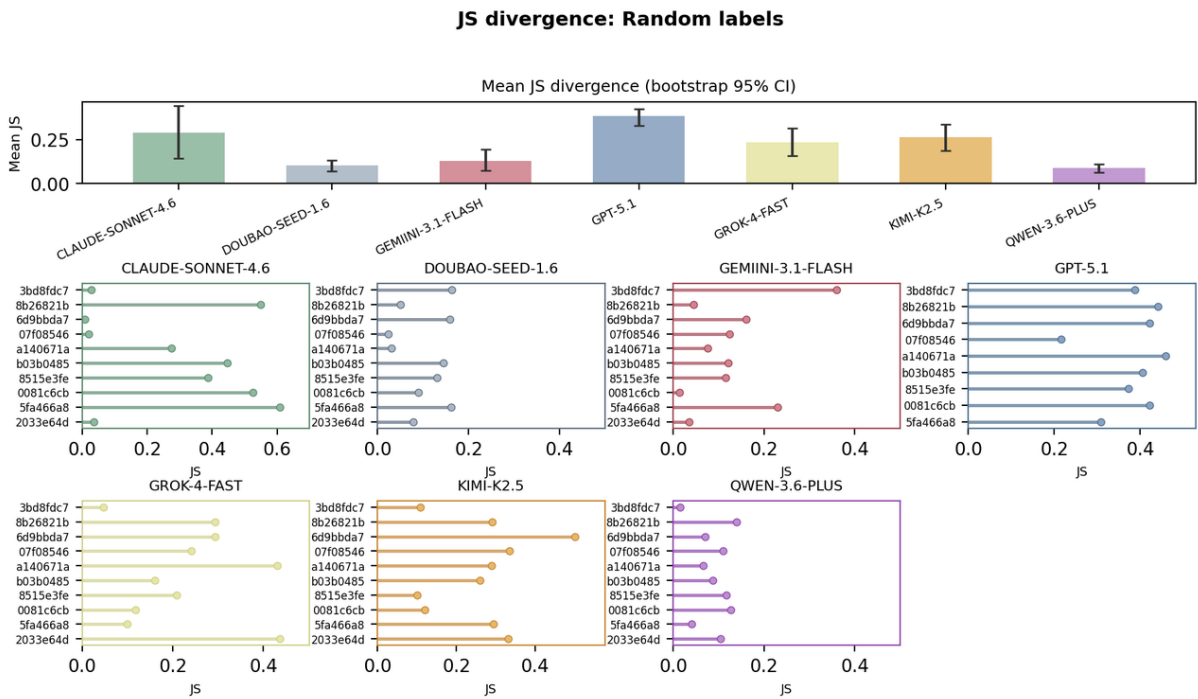


Figure 18: JS divergence: Random labels