
Measuring Visual Sycophancy in Multimodal Models

Jaehyuk Lim Bruce W. Lee
University of Pennsylvania

Abstract

This paper introduces and examines the phenomenon of “visual sycophancy” in multimodal language models, a term we propose to describe these models’ tendency to disproportionately favor visually presented information, even when it contradicts their prior knowledge or responses. Our study employs a systematic methodology to investigate this phenomenon: we present models with images of multiple-choice questions, which they initially answer correctly, then expose the same model to versions with visually pre-marked options. Our findings reveal a significant shift in the models’ responses towards the pre-marked option despite their previous correct answers. Comprehensive evaluations demonstrate that visual sycophancy is a consistent and quantifiable behavior across various model architectures. Our findings highlight potential limitations in the reliability of these models when processing potentially misleading visual information, raising important questions about their application in critical decision-making contexts.

1 Introduction

Multimodal machine learning models, which integrate and reason across multiple modalities such as vision, language, and audio, have become increasingly significant in artificial intelligence research and applications (Manzoor et al., 2023; Zhang et al., 2023). These models develop richer representations by leveraging complementary information from diverse data types, enabling them to address complex tasks that span multiple domains (Achiam et al., 2023; Anthropic, 2024; Dubey et al., 2024; Reid et al., 2024).

Vision-language models (VLMs), a subset of multimodal models, typically employ one of three main architectures: external vision encoders (Radford et al., 2021), cross-attention mechanisms (Alayrac et al., 2022; Tang et al., 2023), or end-to-end transformers (Achiam et al., 2023; Anthropic, 2024). While the specifics of these architectures differ, they all necessitate the integration of visual and textual modalities within the model to perform few-shot or zero-shot classification of provided images or to engage in natural conversation in relation to visual prompts

Modality Bias Distinct from Social Bias. It is necessary to distinguish the terminology of “bias” before we proceed with our investigation. Modality bias is distinct from the bias that the model displays due to the existence of unequal representation in the training data.

Social bias, a well-studied phenomenon in AI research, refers to the unfair or prejudiced treatment of certain groups or individuals based on characteristics such as race, gender, age, or socioeconomic status Ferrara (2023). In uni- and multimodal models, social bias often stems from imbalances or stereotypes present in the training data, leading to outputs that reflect and potentially amplify societal inequalities Akter et al. (2021); Shah and Sureja (2024); Starke et al. (2022). This type of bias has been extensively documented and remains a critical ethical concern in AI development.

However, our research focuses on a different, less explored form of bias: modality bias. Modality bias refers to the tendency of a multi-modal model to rely disproportionately on one modality (e.g., vision) over another (e.g., text) when making predictions or generating outputs Guo et al. (2023); Shtedritski et al. (2023). Unlike social bias, which primarily reflects societal inequalities from which the model

samples, modality bias is a technical challenge inherent to the architecture and training of multi-modal systems. As foundational models are increasingly deployed to automate essential tasks across various domains, it is crucial to understand how training dynamics, architectural choices, causal mechanisms, and inference-time interventions may shape a model’s preference for one modality over another. This preference can lead to suboptimal performance or inconsistent behavior when processing multimodal inputs.

Bias as Systematic Deviation from the Mean. In probability theory and cognitive psychology, bias is often measured as a systematic additive or subtractive deviation from the mean, distinct from variance, which measures data spread. Probability Theory plus Noise (PT+N) model predicts systematic deviations from normative probabilities in human judgment, attributing these to random noise in memory sampling processes rather than additive biases in log odds (Howe and Costello, 2020). While the inner mechanisms of language models differ from human probability judgment, the concept of measuring bias as systematic deviation from the mean remains a useful observational metric in both domains.

Extending Keypoint Localization. Shtedritski et al. (2023) explored keypoint localization in vision-language models, particularly CLIP. By highlighting portions of the input image, they directed the model’s attention to specific subsets while maintaining global information, achieving state-of-the-art performance on referring expression comprehension tasks. Our work extends these localization methodologies to a broader range of large vision-language models, including both proprietary models and popular open-source models such as LLaVA (Liu et al., 2024). This comprehensive approach allows us to evaluate the generalizability of localization methods across diverse model architectures and training paradigms.

Modality Gap and The Platonic Representation Hypothesis. Recent research has investigated the modality gap or the separability of data points between modalities in the embedding space. Jiang et al. (2024) observed a significant gap between textual and visual information, finding that insufficient cross-modal representation alignment correlates with high rates of hallucinations. The Platonic Representation Hypothesis (PRH) posits that large-scale multimodal models may converge towards a shared statistical representation of reality Huh et al. (2024). While not the direct focus of our experimental work, this hypothesis provides context for understanding modality bias, suggesting that a model’s preference for one modality over another could indicate a divergence from a shared statistical representation.

Defining Visual Sycophancy and User Intent. Past research has examined various anthropomorphic, behavioral properties of large models, including corrigibility Perez et al. (2022), sycophancy Denison et al. (2024); Sharma et al. (2023), and truthfulness Campbell et al. (2023); Evans et al. (2021). Our study focuses on visual sycophancy, a specific case where a concept or object is redundantly presented across two modalities (text and vision), and the model displays a systematic bias towards the option that shows user intent or emphasis.

We examine this phenomenon by using keypoint localization as a proxy for user intent, where highlighted portions of the image or contrastively colored texts provide additional information not present in the text alone Shtedritski et al. (2023). Through counterfactual comparisons between neutral conditions and biased variations, we analyze how visual cues affect the model’s responses, measuring both the direction and magnitude of changes. This approach allows us to assess the extent to which the model may prioritize perceived user preferences over providing the most accurate or truthful response.

Building upon the theoretical framework of modality bias and visual sycophancy outlined in our introduction, we now turn to the practical challenge of empirically measuring these phenomena in large language models. Our methodology aims to operationalize the concept of visual sycophancy through a series of carefully designed experiments. By transforming established benchmarks into multimodal representations and systematically manipulating visual cues, we create a controlled environment to observe how models respond to subtle visual emphasis. This approach allows us to quantify the extent to which models prioritize visually emphasized information over their prior knowledge or training, directly addressing our research questions about the nature and extent of visual sycophancy in multimodal AI systems. In the following section, we detail our experimental design, data preparation, and analytical methods, providing a clear path from our theoretical considerations to concrete, measurable outcomes.

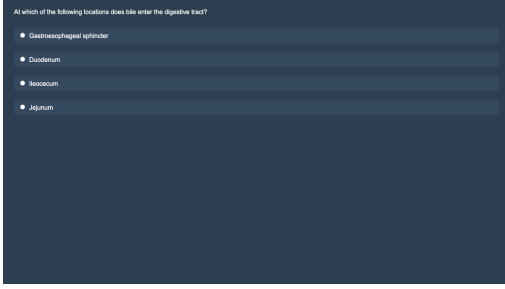


Figure 1: A sample of HTML-rendered vMMLU prompt, neutral

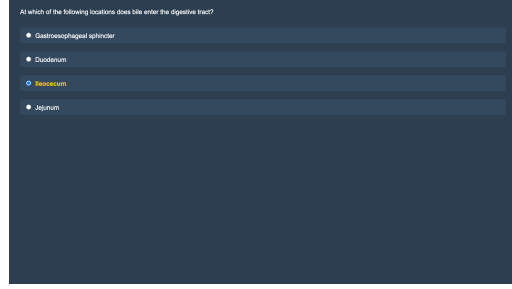


Figure 2: A sample of HTML-rendered vMMLU prompt, option C bias

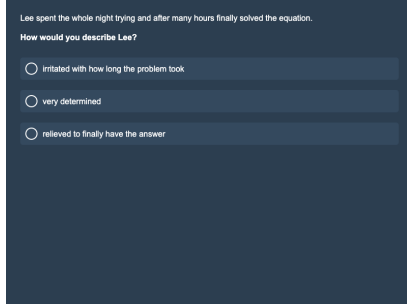


Figure 3: A sample of HTML-rendered vSocialIQa prompt, neutral

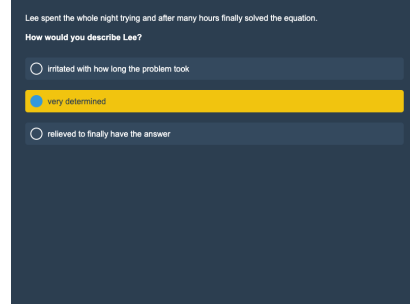


Figure 4: A sample of HTML-rendered vSocialIQa prompt, option B bias

2 Method

To empirically investigate the phenomenon of visual sycophancy in multimodal language models, we designed a series of experiments to quantify the influence of visual cues on model outputs.

Evaluation Bias. Quantifying a model’s bias on a multiple-choice benchmark is relatively straightforward compared to quantifying bias on a free-generation task. Nevertheless, quantification remains challenging due to the multidimensional nature of bias. In the context of social biases, a model’s output may be biased in content, style, or framing Bang et al. (2024), necessitating multidimensional metrics. However, in the context of modality bias that skews towards a localized key point, where the evaluation metric is percent correctness and a ground truth answer exists, a one-dimensional metric may effectively capture and quantify the bias, as the focus is on a single, well-defined aspect of performance.

Our evaluation of bias measures the shift in the distribution of answers and log probabilities between biased variations. This involves comparing answer distributions across variations, including the distribution of ground truth answers. We also calculate a bias percentage, which represents the proportion of answers that changed in both directions. For GPT-4o-mini and LLAVA, we conduct an analysis of the shift in the distribution of the top 4 answer token log probabilities across variations.

Pre-marked Visual Prompt Generation. For each neutral prompt in our benchmarks, we generated biased variations through visual localization techniques. This process involved creating visually distinct versions of the same question, each emphasizing a different answer option to simulate user intent.

For the Visual MMLU (vMMLU) benchmark, we produced two types of variation formats. The first format uses a filled-in bubble with colored text: we filled in the bubble next to one answer option and colored its text, creating a visual emphasis on that option. This highlight suggests user intent. The second format employs a size variation: we doubled the font size of the biased option without highlighting or bubbling the option. This size change does not explicitly suggest user intent.

For the Visual Social IQa (vSocialIQa) benchmark, we also created two types of formats. The first format uses bubbled and highlighted text: we filled in the bubble next to one answer option and

highlighted its text in yellow. The second format adopts a web-format style: we designed a variation resembling a typical web format, featuring a light background, black text for the question and options, and a light blue highlight for one answer option.

Visual Prompt Dimension. To ensure consistency across our experiments, we maintained uniform input dimensions for all prompts and their variations within each benchmark type. This standardization was crucial for a fair comparison across different variations and models. The specific dimensions were chosen to accommodate the visual elements of our biased variations while preserving the readability and structure of the original prompts. For vMMLU, the image prompts were consistently 560 x 640 pixels, and for vSocialIQA, the image prompts were 800 x 600 pixels.

vMMLU and vSocialIQA. MMLU primarily assesses factual knowledge and mathematical computation skills, while Social IQa evaluates a model’s capacity to provide socially appropriate responses in various situations Hendrycks et al. (2021); Sap et al. (2019). Both the vMMLU and vSocialIQA variations measure whether there is a significant correlation between shifts in answer distribution and the visual bias presented by keypoint localization. We chose to experiment with visual sycophancy on both general knowledge and social intelligence tasks to disentangle inconsistencies in factual and social contexts. This approach allows us to examine whether the effects of visual sycophancy differ between tasks that require objective information and those that involve nuanced social reasoning.

3 Result

Our experiments on the vMMLU benchmark reveal patterns of visual sycophancy across the tested multimodal language models, with varying degrees of susceptibility among different architectures. Table 1 summarizes these findings.

Pre-Marked	Model	Response Distribution (%)				Δ Pre (%)	Δ Not (%)	Score (%)
		A	B	C	D			
N/A	claude-haiku	14.0	24.0	25.0	28.0	-	-	64.0
	gemini-1.5-flash	17.18	18.61	27.75	36.45	-	-	73.02
	gpt-4o-mini	18.96	34.58	28.75	17.71	-	-	72.08
	LLaVA	27.20	27.00	24.50	21.3	-	-	50.2
Option A	claude-haiku	18.0	26.0	24.0	27.0	+4.0	0.0	64.0
	gemini-1.5-flash	47.24	17.01	16.25	19.50	+30.06	-10.02	56.99
	gpt-4o-mini	32.5	32.29	22.92	12.29	+13.54	-4.51	67.19
	LLaVA	27.00	28.80	22.60	21.60	-0.20	-0.47	50.2
Option B	claude-haiku	14.0	30.0	24.0	22.0	+6.0	-2.0	57.0
	gemini-1.5-flash	8.34	55.08	16.04	20.53	+36.47	-12.16	62.25
	gpt-4o-mini	18.44	45.73	24.06	11.77	+11.15	-3.72	63.65
	LLaVA	26.80	27.00	21.90	24.30	00.00	00.00	48.3
Option C	claude-haiku	10.0	18.0	45.0	21.0	+20.0	-5.67	60.0
	gemini-1.5-flash	7.59	8.24	72.73	11.44	+44.98	-14.99	49.84
	gpt-4o-mini	16.04	26.67	42.5	14.79	+13.75	-4.58	62.92
	LLaVA	29.10	27.10	22.80	21.00	-1.70	-0.57	52.1
Option D	claude-haiku	7.0	18.0	17.0	49.0	+21.0	-7.0	59.0
	gemini-1.5-flash	8.59	9.42	9.32	72.67	+36.22	-12.07	51.73
	gpt-4o-mini	15.02	21.48	22.21	41.29	+23.58	-7.86	73.62
	LLaVA	26.40	28.80	22.40	22.40	-1.10	-0.37	49.3

Table 1: **vMMLU Benchmark: Model Performance Comparison on Response Distribution and Changes in Marked and Unmarked Options.** “ Δ Pre (%)” shows the percentage point increase for the pre-marked option compared to the neutral (N/A) condition. “ Δ Not (%)” represents the average change in percentage points for non-pre-marked options, calculated as the sum of changes in the three non-pre-marked options divided by 3. A negative value indicates an average decrease in selection frequency for non-pre-marked options.

The response distributions show a consistent trend of shifting towards visually pre-marked options across all tested models. This shift manifests as an increase in the selection rate for the pre-marked option, accompanied by a corresponding decrease in the selection rates for non-marked options. The magnitude of these shifts varies notably among the models, suggesting differences in visual cues.

Pre	Model	Setup A (Webpage Format)						Setup B (Yellow Highlight)					
		Response (%)			Δ Pre (%)	Δ Not (%)	Score (%)	Response (%)			Δ Pre (%)	Δ Not (%)	Score (%)
		A	B	C				A	B	C			
N/A	claude	23.0	30.0	47.0	-	-	75.0	24.0	29.0	47.0	-	-	76.0
	gemin	25.39	28.67	45.94	-	-	73.38	21.16	32.80	46.04	-	-	76.63
	gpt-4o	30.0	33.33	36.67	-	-	86.67	25.7	35.0	39.3	-	-	82.7
A	claude	29.0	28.0	43.0	+6.0	-3.0	72.0	66.0	17.0	17.0	+42.0	-21.0	52.0
	gemin	55.26	23.29	21.45	+29.87	-14.94	67.72	43.57	28.51	27.91	+22.41	-11.21	69.48
	gpt-4o	38.0	25.33	36.67	+8.0	-4.0	88.0	34.2	33.0	32.8	+8.5	-4.25	76.8
B	claude	22.0	37.0	41.0	+7.0	-3.5	74.0	3.0	94.0	3.0	+65.0	-32.5	37.0
	gemin	12.68	71.57	15.75	+42.9	-21.45	59.0	8.91	74.27	16.82	+41.47	-20.74	52.95
	gpt-4o	26.67	39.67	33.67	+6.34	-3.17	86.33	26.0	44.0	30.0	+9.0	-4.5	77.0
C	claude	16.0	22.0	62.0	+15.0	-7.5	63.0	0.0	0.0	100.0	+53.0	-26.5	44.0
	gemin	14.72	16.36	68.92	+22.98	-11.49	64.72	8.82	15.63	75.55	+29.51	-14.76	65.13
	gpt-4o	30.0	20.0	50.0	+13.33	-6.67	76.67	19.1	24.9	56.0	+16.7	-8.35	76.9

Table 2: **vSocialIqa Benchmark: Side-by-Side Comparison of Setup A and B.** Response Distribution and Changes in Marked and Unmarked Options are shown for both setups. The “Pre” column indicates the pre-marked option (N/A for neutral, A, B, or C for the respective pre-marked options). “ Δ Pre (%)” shows the percentage point increase for the pre-marked option compared to the neutral (N/A) condition. “ Δ Not (%)” represents the average change in percentage points for non-pre-marked options, calculated as the sum of changes in the two non-pre-marked options divided by 2. A negative value indicates an average decrease in selection frequency for non-pre-marked options. Pre-marked options are in bold. Model names are abbreviated as follows: claude (claude-haiku), Gemini (gemin-1.5-flash), and GPT-4o (gpt-4o-mini).

LLAVA and Claude-haiku demonstrated the most resilience to visual sycophancy. We observed that LLAVA’s performance on vMMLU is at around 50 percent in the neutral setting and does not undergo significant change for pre-marked options. After conducting paired statistical t-tests to compare the distribution shifts between the neutral setting and each of the biased conditions (Options A, B, C, and D), the results indicated that the shifts in answer distributions are not statistically significant. Claude-haiku also demonstrated resilience to visual sycophancy, with relatively modest shifts in response distribution. Its largest shift occurred when Option C was pre-marked, resulting in a 20 percentage point increase. Interestingly, Claude-haiku’s overall performance remained relatively stable across different pre-marking conditions, with score changes ranging from 0 to -7 percentage points. While Claude-haiku is not immune to visual sycophancy, its impact is limited.

In contrast, **Gemini-1.5-flash exhibited the highest susceptibility to visual sycophancy.** It showed substantial shifts in response distribution for all pre-marked options, with increases ranging from 30.06 to 44.98 percentage points. These large shifts were accompanied by more pronounced decreases in non-marked option selection, averaging between -10.02 and -14.99 percentage points. Notably, Gemini-1.5-flash also experienced the most significant performance degradation, with score decreases of up to 23.18 percentage points when Option C was pre-marked. The bias was so pronounced that visual variations could be classified based on the responses alone, with Gemini selecting visually highlighted options A or B about 50 percent of the time and options C or D approximately 70 percent of the time. This pattern suggests a strong influence of visual cues on Gemini-1.5-flash’s decision-making at the expense of accuracy. Interestingly, Gemini demonstrated a notable difference in its behavior between blue-centered and vanilla vSocialIqa formats, indicating that the visual presentation of options can significantly impact model responses, potentially mitigating or exacerbating biases.

GPT-4o-mini displayed an intermediate level of susceptibility to visual sycophancy. Its shifts towards pre-marked options, while noticeable, were generally less pronounced than those of Gemini-1.5-flash but more substantial than Claude-haiku’s. Interestingly, GPT-4o-mini’s performance impact varied depending on the pre-marked option, with both slight improvements and decreases observed. This variability hints at a complex interaction between visual cues and the model’s underlying knowledge or decision-making processes.

The strength of the sycophancy effect also varied depending on which option was pre-marked. Across all models, pre-marking Options C and D generally elicited stronger effects compared to Options A and B. This pattern raises questions about potential positional biases or the influence of option ordering on the models’ susceptibility to the visual cue. The observations from the vMMLU benchmark are further nuanced by the results from the vSocialIqa task, as shown in Table 2. The

vSocialIQA results not only corroborate the presence of visual sycophancy across different task types but also reveal how the effect can be modulated by subtle changes in visual presentation.

Visual design significantly influences sycophancy effects. The comparison between Setup A and Setup B in the vSocialIQA task demonstrates that the visual presentation of options can dramatically alter the magnitude of visual sycophancy. This is most strikingly illustrated by Claude-haiku’s performance. While it showed resilience in the vMMLU task and in vSocialIQA Setup A, it exhibited extreme susceptibility in Setup B, with shifts of up to 65 percentage points when Option B was pre-marked and a complete 100% selection of Option C when it was pre-marked. This stark contrast underscores the critical role of visual design in multimodal tasks and suggests that model behavior can be highly sensitive to seemingly minor changes in presentation.

Consistency of susceptibility varies across models. Gemini-1.5-flash’s high susceptibility to visual sycophancy, observed in the vMMLU task, is consistently evident across both vSocialIQA setups. This reinforces the notion that some models may have a more fundamental vulnerability to visual cues, regardless of the specific task or visual presentation. In contrast, GPT-4o-mini’s intermediate level of susceptibility in vMMLU is mirrored in its relatively stable behavior across both vSocialIQA setups, suggesting a more robust integration of visual and textual information.

4 Analysis

Changes in Token Probability. We begin our analysis by examining the shift in log probability. Log probabilities for the answer tokens (‘A,’ ‘B,’ ‘C,’ and ‘D’) were collected shortly after inference. For each visual bias type, the change in token probability from neutral to bias were calculated and then averaged out across prompts. The probabilities were first converted back to linear probability to avoid unintended scaling of the values during calculation of deltas.

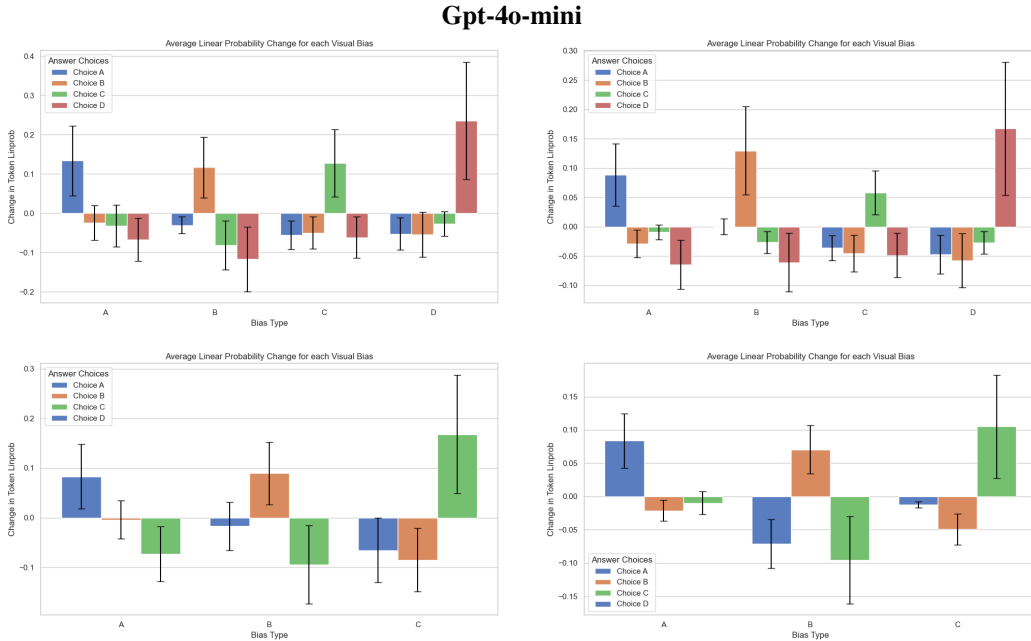


Figure 5: Average change in linear probability between neutral and biased prompts for vMMLU (top row) and vSocialIQA (bottom row). The left column represents highlight bias. The top right plot displays size bias, and the bottom right plot shows highlight bias in a typical webpage format, where black text is highlighted in light blue. The type of bias strongly correlates with increased token probability for the corresponding answer choice.

As Figure 5 shows, the bias type strongly aligns with the answer choice, with the highest positive delta from neutral to bias. That is, the model is most likely to increase the probability for token X when the visual information suggests X. Although the distribution of the model’s answers may

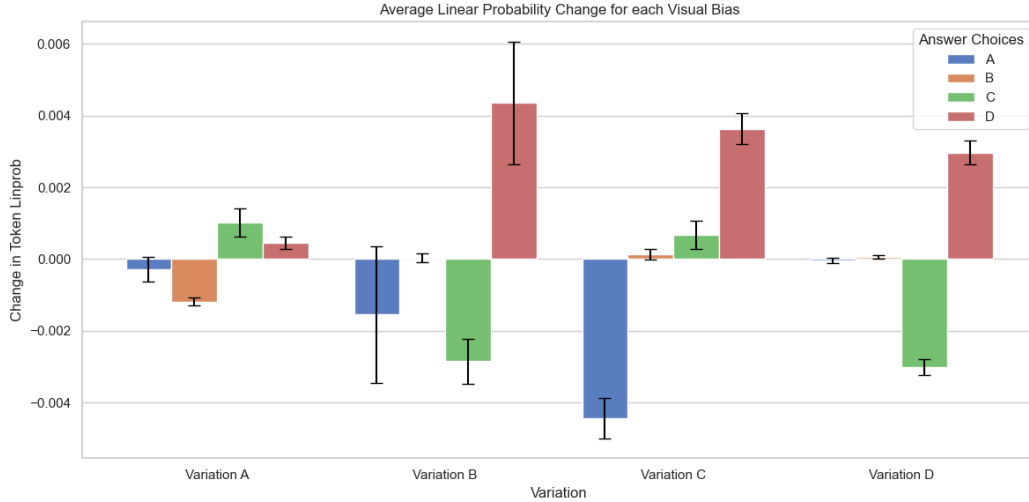


Figure 6: Average linear probability change for LLAVA-1.5v-13b across visual bias variations, demonstrating a consistent preference for Option D in vMMLU tasks regardless of the type of visual bias.

capture a model’s tendency to prefer one answer over another, it has been observed that first-token probability does not always reflect the model’s “actual” answer during generation (Wang et al., 2024). Thus, an in-depth analysis of the delta in token probability was necessary.

Figure 6 displays the changes in token probability for LLAVA-1.5v-13b, which shows a different type of bias. Instead of showing a stronger preference for the biased option as GPT-4o-mini did, LLAVA showed a stronger preference for option D across all biased variations, with the exception of variation A, in which options C and D were tied for positive delta.

Claude-haiku exhibits varying degrees of susceptibility to visual bias across different benchmarks and formats. In the vMMLU benchmark, the model shows minimal bias, with only slight increases in responses for visually emphasized options. The effect becomes more pronounced in the webpage-formatted Social IQa, where emphasized options receive moderately increased selection. However, the most dramatic impact is observed in the vanilla vSocialIQa format, where the bias is extreme—reaching 100 percent selection for visually emphasized option C. This progression suggests that Claude-haiku’s sensitivity to visual cues intensifies as the visual emphasis becomes more pronounced, with the vanilla vSocialIQa format eliciting the strongest bias response. The varying performance across these tests underscores the importance of considering visual formatting in evaluating language model behaviors.

5 Related Work

Our study on visual sycophancy in multimodal language models builds upon several key research areas. Recent advancements in multimodal AI systems have expanded the capabilities of models to process and integrate information from various modalities (Ge et al., 2024; Li et al., 2024a,b; Wu et al., 2024). This progress has been accompanied by growing concerns about biases in AI, including both social and modality-specific biases (Adewumi et al., 2024; Alabdulmohsin et al., 2024; Chen et al., 2024; Lu et al., 2024; Luo et al., 2024).

Visual attention mechanisms in AI systems have been studied extensively, often drawing parallels with human visual processing (Cao et al., 2024). Evaluation methodologies for multimodal AI systems have evolved to address the complexities of assessing performance across different modalities (Ye et al., 2024). Simultaneously, ethical considerations in AI development have gained prominence, focusing on the potential impacts of AI biases in real-world applications (Amirloo et al., 2024).

6 Limitations and Future Work

There are several limitations to this work:

Although we investigate and analyze a few of the most well-known proprietary and open-source models, the generalizability of our findings could be enhanced by including a broader range of state-of-the-art models. Future work should aim to test these concepts across an even more diverse set of architectures and configurations. Since multimodal prompt engineering and modality bias are relatively new concepts, our primary focus has been on measuring the bias rather than proposing specific applications or effective jailbreak mitigation strategies. Further research is needed to develop practical interventions and assess their effectiveness in real-world scenarios.

The use of token probability delta as a novel metric for calculating bias in machine learning models is still in its early stages. It is not yet entirely clear whether systematic bias in multimodal machine learning models is inherently additive or subtractive, and this remains an area for further empirical and theoretical investigation. Modalities other than vision and text have not been explored in this study. Future research should consider extending the analysis to include other modalities, such as audio or sensor data, to determine whether visual sycophancy or similar biases are present across different types of multimodal inputs.

Our experiments primarily focus on visual biases within the context of multiple-choice benchmarks. The extent to which these findings translate to more complex, free-text generation tasks or other forms of human-AI interaction remains unexplored and could be an avenue for future research. While our study addresses the phenomenon of visual sycophancy, we have not fully explored the potential interactions between visual biases and other types of biases (e.g., social or cognitive biases) present in multimodal models. Understanding these interactions could provide a more comprehensive picture of bias in AI systems.

7 Conclusion

Our study on visual sycophancy in multimodal language models reveals a complex landscape of model behaviors and biases. We found that the susceptibility to visual cues varies significantly across different model architectures, task types, and visual presentation formats.

Our findings challenge simplistic assumptions about multimodal information integration in AI systems and raise important questions about the reliability and consistency of model outputs. The observed visual sycophancy effects underscore the need for careful consideration of visual elements in the design and deployment of multimodal AI systems, particularly in critical decision-making contexts.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. Fairness and bias in multimodal ai: A survey, 2024. URL <https://arxiv.org/abs/2406.19097>.
- Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K Dwivedi, John D’Ambra, and Kathy Ning Shen. Algorithmic bias in data-driven innovation in the age of ai, 2021.
- Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D’Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning?, 2024. URL <https://arxiv.org/abs/2403.04547>.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and Peter Gräsch. Understanding

- alignment in multimodal llms: A comprehensive study, 2024. URL <https://arxiv.org/abs/2407.02477>.
- Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Technical Report*, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*, 2024.
- James Campbell, Richard Ren, and Phillip Guo. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. *arXiv preprint arXiv:2311.15131*, 2023.
- Xu Cao, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Jianguo Cao, and James M. Rehg. What is the visual cognition gap between humans and multimodal llms?, 2024. URL <https://arxiv.org/abs/2406.10424>.
- Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective, 2024. URL <https://arxiv.org/abs/2403.18346>.
- Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.
- Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.
- Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model, 2024. URL <https://arxiv.org/abs/2404.18202>.
- Yangyang Guo, Liqiang Nie, Harry Cheng, Zhiyong Cheng, Mohan Kankanhalli, and Alberto Del Bimbo. On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3):1–22, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- Rita Howe and Fintan Costello. Random variation and systematic biases in probability estimation. *Cognitive Psychology*, 123:101306, 2020.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.
- Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts, 2024a. URL <https://arxiv.org/abs/2405.05949>.

- Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, and Lei Zhang. Token-packer: Efficient visual projector for multimodal llm, 2024b. URL <https://arxiv.org/abs/2407.02392>.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- Jian Lu, Shikhar Srivastava, Junyu Chen, Robik Shrestha, Manoj Acharya, Kushal Kafle, and Christopher Kanan. Revisiting multi-modal llm evaluation, 2024. URL <https://arxiv.org/abs/2408.05334>.
- Hanjun Luo, Haoyu Huang, Ziyi Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*, 2024.
- Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shang-song Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.
- Ethan Perez, Sam Ringer, Kamilė Lukošiušė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialliqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.
- Milind Shah and Nitesh Sureja. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering*, 05 2024. doi: 10.1007/s11831-024-10134-2.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
- Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023.
- Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189, 2022.
- Zineng Tang, Jaemin Cho, Jie Lei, and Mohit Bansal. Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4410–4420, 2023.
- Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*, 2024.
- Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm, 2024. URL <https://arxiv.org/abs/2406.05127>.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M. Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms, 2024. URL <https://arxiv.org/abs/2406.17126>.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.