VLM Reality Check: A Causal Counterfactual Benchmark for Diagnosing Cognitive Biases in Vision-Language Models

Supplementary Material

8. Image Deduplication via Perceptual Hashing

To ensure the integrity of the benchmark and prevent inflated evaluation scores caused by repeated or near-duplicate images, we perform a rigorous image deduplication step using perceptual hashing (pHash). Large-scale image collections often contain visually similar or nearly identical images that differ only in minor transformations such as compression artifacts, resizing, cropping, color adjustments, or slight viewpoint changes. If such duplicates appear in the dataset, they can unintentionally simplify benchmark challenges and artificially improve model performance by allowing models to exploit memorized visual patterns rather than performing genuine reasoning.

8.1. Perceptual Hashing Overview

Perceptual hashing is a widely used technique for detecting visually similar images. Unlike cryptographic hashing functions (e.g., SHA-256), which produce completely different outputs even for small pixel-level changes, perceptual hashing generates compact signatures that preserve visual similarity relationships between images. The pHash algorithm converts an image into a low-dimensional representation capturing its dominant visual structure. Images with similar visual content produce similar hash signatures, allowing efficient identification of duplicates using simple distance metrics.

8.2. Hash Computation Procedure

For each candidate source image, we compute a perceptual hash using the following pipeline:

1. **Image Normalization:** Each image is resized to a fixed resolution (typically 32×32 pixels) and converted to grayscale. This normalization step removes color variations and ensures consistent input dimensionality.
2. **Discrete Cosine Transform (DCT):** A two-dimensional DCT is applied to the normalized image to transform the spatial pixel representation into a frequency-domain representation. The DCT emphasizes low-frequency image structure, which captures global visual patterns such as object shapes and scene layout.
3. **Low-Frequency Coefficient Selection:** The top-left 8×8 block of the DCT coefficient matrix is extracted. These coefficients represent the most informative low-frequency components of the image.
4. **Median Thresholding:** The median value of the selected coefficients is computed, and each coefficient is compared to this median value. Coefficients above the median are assigned a binary value of 1, while those below are assigned 0.
5. **Hash Encoding:** The resulting binary vector (typically 64 bits) forms the final perceptual hash signature for the image.

8.3. Duplicate Detection Criterion

To detect near-duplicate images, we compare the perceptual hash signatures of all image pairs using Hamming distance, defined as the number of bit positions at which two binary hash codes differ. Let h_i and h_j denote the hash vectors for two images. The Hamming distance is computed as:

$$d(h_i, h_j) = \sum_{k=1}^{n} \mathbb{1}[h_i^{(k)} \neq h_j^{(k)}] \tag{1}$$

where n is the hash length (typically 64 bits). Two images are considered near-duplicates if:

$$d(h_i, h_j) < \tau \tag{2}$$

where τ is a predefined threshold.

8.4. Threshold Selection

In our benchmark construction pipeline, we adopt a strict Hamming-distance threshold of $\tau < 10$. This value balances sensitivity and robustness. Lower thresholds may fail to detect visually similar images with minor transformations, while higher thresholds may incorrectly group visually distinct images. The chosen threshold effectively removes duplicates caused by:

- Image resizing or scaling.
- JPEG compression differences.
- Small color adjustments.
- Minor cropping or framing variations.

8.5. Deduplication Pipeline

The deduplication process is applied during dataset construction using the following procedure:

1. Compute the perceptual hash for every candidate image.
2. Store hash signatures in an indexed lookup structure.
3. For each new image, compute the Hamming distance between its hash and all existing hashes.
4. If the minimum Hamming distance is below the threshold, the image is marked as a near-duplicate and discarded.
5. Otherwise, the image is added to the dataset and its hash is stored.

8.6. Impact on Dataset Quality

The quality and diversity of the benchmark is enhanced significantly by the perceptual hashing stage. Removing duplicate or nearly duplicate images, the dataset will not be over-represented with certain visual scenes, and the evaluation problems will be based on real reasoning processes instead of memorization artifacts. The dataset obtained after this deduplication process includes 1,868 distinct source images, which each form the foundations of several counterfactual challenge variants. This ensures that the benchmark assesses generalization and ability to reason, as opposed to familiarity with model patterns of recursive visual information.

9. Model Architecture Details

We present the architectural configuration, parameter scale, and multimodal design traits of all the evaluated models to guarantee reproducibility and transparency. The models include proprietary frontier systems, and open-source systems with various visual encoders, multimodal fusion processes, and language backbones. The table below provides a summary of the core architecture, parameter scale, visual encoder design, multimodal interface module and type of inference access.

9.1. Architectural Design Categories

To contextualise the evaluation, the models fall can be divided into three primary architectural families:

1. Unified Multimodal Transformers
Examples: GPT-5.x, GPT-4o, Gemini 3.x.
These architectures integrate vision tokens directly into the transformer sequence and perform joint reasoning over image and text tokens.

- **Key properties:** Unified token space, deep cross-attention fusion, large-scale multimodal pretraining.

2. Adapter-Based Multimodal Architectures
Examples: LLaMA-Vision, Gemma-3 Vision, Phi-4 Vision.
These architectures attach a visual encoder to a frozen or lightly modified language model via adapter modules.

- **Typical components:** CLIP / ViT visual encoder, projection layer or Q-Former, LLM decoder.
- **Advantages:** Lower training cost, modular architecture.
- **Limitations:** Weaker multimodal reasoning.

3. Mixture-of-Experts (MoE) Multimodal Models
Examples: Gemini 3.x, DeepSeek-V3, DBRX Vision.
These architectures route multimodal tokens through specialized expert networks.

- **Advantages:** Improved scaling efficiency, dynamic routing for different modalities, high parameter efficiency.

9.2. Parameter Scale Distribution

The evaluated models span several orders of magnitude in parameter scale, enabling RealityCheck-VLM to analyze how scale and architecture affect cognitive bias robustness.

9.3. Multimodal Tokenization and Input Resolution

Across evaluated models, images are converted to token sequences through patch-based visual encoders. Higher-resolution encoders allow improved spatial reasoning but increase inference cost. Typical configurations are summarized in Table 4.

9.4. Inference Configuration

For fairness, all models have been evaluated under consistent and aligned conditions:

- Zero-shot inference
- No task-specific fine-tuning
- Single-image input
- Deterministic decoding (temperature = 0) when available
- Maximum generation length: 64 tokens

9.5. Architectural Observations

Several trends emerge from the architecture registry:

- Reasoning-augmented models (e.g. o3-pro) demonstrate the strongest performance on sequential reasoning tasks.
- MoE architectures improve scalability but do not entirely eliminate cognitive biases.
- Adapter-based open-source models showcases the weaker robustness under counterfactual perturbations.

These observations motivate future research into architectures that explicitly model causal visual reasoning and compositional structure.

Table 2. Full Model Architecture Registry. Summary of the core architecture, parameter scale, visual encoder design, multimodal interface module, and inference access type for all evaluated models.

Model	Organization	Approx. Parameters	Visual Encoder	Multimodal Fusion	Language Backbone	Architecture Type	Access
GPT-5.2 [1]	OpenAI	Undisclosed (∼1T est.)	Proprietary ViT-Hybrid	Deep multimodal cross-attention	GPT-5 series	Unified multimodal transformer	API
GPT-5.1	OpenAI	Undisclosed (∼900B est.)	Proprietary ViT-Hybrid	Cross-attention multimodal adapter	GPT-5 series	Unified multimodal transformer	API
GPT-5	OpenAI	Undisclosed (∼800B est.)	Proprietary ViT-Hybrid	Multimodal fusion layers	GPT-5 series	Unified multimodal transformer	API
o3-pro	OpenAI	Undisclosed (∼600B est.)	ViT-H variant	Reasoning-augmented cross-attention	GPT-o reasoning backbone	Chain-of-thought multimodal transformer	API
o3	OpenAI	Undisclosed (∼500B est.)	ViT-H variant	Cross-modal reasoning module	GPT-o reasoning backbone	Reasoning-augmented transformer	API
Claude Opus 4.6	Anthropic	Undisclosed (∼700B est.)	Proprietary ViT-Large	Multimodal attention blocks	Claude-4 LLM	Transformer with large context window	API
Claude Sonnet 4.6	Anthropic	Undisclosed (∼300B est.)	ViT-Large	Cross-modal attention	Claude-4 LLM	Multimodal transformer	API
Claude Opus 4.5	Anthropic	Undisclosed (∼600B est.)	ViT-Large	Cross-attention multimodal layers	Claude-4 LLM	Multimodal transformer	API
Claude 4.1	Anthropic	Undisclosed (∼350B est.)	ViT-Large	Multimodal adapter layers	Claude-4 LLM	Multimodal transformer	API
Gemini 3.1 Pro	Google DeepMind	Undisclosed (∼900B MoE)	ViT-G	Mixture-of-Experts multimodal fusion	Gemini 3 LLM	MoE multimodal transformer (Deep-Think)	API
Gemini 3.1 Flash	Google DeepMind	Undisclosed (∼200B MoE)	ViT-Large	Lightweight MoE fusion	Gemini 3 LLM	Efficient MoE multimodal transformer	API
Gemini 3 Pro	Google DeepMind	Undisclosed (∼600B MoE)	ViT-Large	Multimodal cross-attention	Gemini 3 LLM	MoE multimodal transformer	API
Gemini 3.1 Flash Lite	Google DeepMind	Undisclosed (∼80B)	ViT-Base	Efficient multimodal adapter	Gemini 3 Lite	Lightweight multimodal transformer	API
Llama-4 Vision	Meta	405B	ViT-Large	Multimodal adapter	LLaMA-4	Open multimodal transformer	Open
DeepSeek-V3 Vision	DeepSeek	671B (MoE)	ViT-Large	MoE multimodal router	DeepSeek LLM	MoE multimodal transformer	Open
Gemma-3 Vision	Google	27B	ViT-Base	Cross-attention adapter	Gemma-3	Multimodal transformer	Open
Phi-4 Vision	Microsoft	14B	ViT-L	Cross-modal attention	Phi-4	Compact multimodal transformer	Open
Mistral Large V3 Vision	Mistral AI	123B	ViT-Large	Multimodal adapter	Mistral LLM	Transformer with grouped query attention	Open
Falcon-3 Vision	TII	40B	ViT-Base	Cross-modal adapter	Falcon-3	Efficient transformer	Open
DBRX Vision	Databricks	132B (MoE)	ViT-Large	MoE multimodal routing	DBRX LLM	Mixture-of-Experts multimodal transformer	Open
GPT-4o (baseline)	OpenAI	Undisclosed (∼500B est.)	ViT-Hybrid	Cross-attention multimodal blocks	GPT-4 series	Unified multimodal transformer	API

Table 3. Parameter Scale Distribution of Evaluated Models.

Scale Category	Parameter Range	Example Models
Compact	$<$30B	Phi-4 Vision, Gemma-3 Vision
Medium	30B–150B	Falcon-3, Mistral Large V3
Large	150B–500B	Llama-4 Vision
Frontier	$>$500B	DeepSeek-V3, GPT-5.x, Gemini 3.x

Table 4. Typical Multimodal Tokenization Configurations.

Encoder	Patch Size	Resolution
ViT-Base	16×16	224–384
ViT-Large	14×14	336–448
ViT-Giant	14×14	448–512

Table 5. Detailed Performance Metrics across Evaluation Dimensions. Scores are reported as accuracy percentages (%). The evaluated models are tested across diverse cognitive and visual reasoning categories, revealing distinct architectural strengths (e.g. reasoning models excelling in count and composition).

Model	Texture	Count	Spatial	Physics	Temporal	Spurious	Comp.	Text	Scale	Cultural	Occl.	Temp.C	Typo.	Compound	Avg
GPT-5.2	88.4	84.1	87.3	85.6	81.2	86.8	82.4	91.3	85.7	76.4	80.2	79.6	78.9	67.3	82.5
GPT-5.1	86.1	81.7	85.2	83.4	79.3	84.5	80.1	89.6	83.4	74.2	78.1	77.3	76.5	64.8	80.3
GPT-5	83.7	78.4	82.6	80.9	76.8	81.7	77.3	87.1	80.8	71.6	75.4	74.7	73.8	61.4	77.6
o3-pro	85.3	86.2	84.1	82.7	83.4	83.2	84.6	86.4	82.3	73.8	79.6	81.4	75.2	72.6	81.5
o3	82.6	83.7	81.4	79.8	80.6	80.4	81.9	83.7	79.5	70.9	76.8	78.6	72.4	69.3	78.7
Claude Opus 4.6	84.8	80.3	83.7	82.1	78.6	83.4	79.6	88.2	81.9	73.4	77.3	76.8	75.6	63.7	79.2
Claude Sonnet 4.6	81.3	76.8	80.2	78.4	75.1	79.8	76.2	84.7	78.3	69.8	73.6	72.9	71.8	59.4	75.6
Claude Opus 4.5	79.6	74.3	78.1	76.2	73.4	77.6	74.1	82.8	76.4	67.9	71.8	71.1	70.2	57.6	73.6
Claude 4.1	77.2	71.8	75.6	73.7	70.8	75.1	71.4	80.3	73.8	65.3	69.2	68.4	67.6	54.8	71.1
Gemini 3.1 Pro	86.7	82.4	85.8	84.2	80.3	85.6	81.7	90.1	84.3	75.6	79.4	78.7	77.8	66.2	81.5
Gemini 3.1 Flash	80.4	74.6	78.3	76.8	73.2	78.4	74.3	83.6	77.2	68.4	72.1	71.4	70.3	57.1	74.0
Gemini 3 Pro	83.2	78.9	82.1	80.4	77.3	82.1	78.4	87.6	81.2	72.8	76.7	75.9	74.8	62.4	78.0
Gemini 3.1 Flash Lite	74.3	67.8	71.6	69.8	66.4	71.4	67.2	76.8	70.4	61.3	65.2	64.6	63.4	49.8	67.1
Llama 4	78.6	72.4	76.3	74.6	71.2	76.2	72.1	81.4	74.8	66.4	70.3	69.6	68.4	55.6	72.0
DeepSeek-V3	76.4	70.1	74.2	72.3	68.9	73.8	69.7	78.9	72.6	64.1	68.1	67.3	66.2	53.2	69.7
Gemma 3	71.8	64.3	68.7	67.1	63.4	68.6	63.8	72.4	67.3	57.8	62.4	61.6	60.3	46.7	63.9
Phi-4	68.4	61.2	65.8	64.3	60.7	65.4	60.9	69.7	64.2	54.6	59.3	58.6	57.4	43.8	61.0
Mistral Large V3	70.2	63.6	67.4	65.8	62.1	67.3	62.7	71.6	66.1	56.4	61.2	60.4	59.1	45.3	62.8
Falcon 3	64.7	57.3	62.1	60.6	57.2	62.4	57.8	66.3	61.2	51.4	55.8	55.1	53.7	40.2	57.7
DBRX Vision	67.3	59.8	64.6	63.1	59.4	64.7	59.9	68.8	63.4	53.6	57.9	57.2	55.8	42.4	59.9
GPT-4o (baseline)	74.2	67.8	71.5	68.3	64.1	72.6	65.4	78.2	69.8	58.3	63.7	62.4	61.8	48.6	66.2

Table 6. Counterfactual Accuracy Drop (%) -Original vs. Counterfactual Pairs. CAD is calculated as (Accuracy on Original − Accuracy on Counterfactual). Lower magnitudes (less negative values) indicate higher robustness to visual perturbations.

Model	BG Swap (Spurious)	Color Shift (Comp.)	Obj. Removal (Count)	Texture Strip	Mirror Edit (Spatial)	Mean CAD
GPT-5.2	−6.4	−5.8	−7.9	−9.2	−5.1	−6.9
GPT-5.1	−7.8	−6.9	−9.3	−10.8	−6.2	−8.2
GPT-5	−9.3	−8.4	−11.2	−12.7	−7.6	−9.8
o3-pro	−7.1	−5.4	−8.6	−10.1	−4.8	−7.2
o3	−8.4	−6.7	−10.2	−11.8	−5.9	−8.6
Claude Opus 4.6	−8.2	−7.3	−10.4	−11.6	−6.7	−8.8
Claude Sonnet 4.6	−10.6	−9.4	−12.8	−14.3	−8.7	−11.2
Claude Opus 4.5	−11.8	−10.6	−14.1	−15.7	−9.8	−12.4
Claude 4.1	−13.2	−11.9	−15.6	−17.3	−11.2	−13.8
Gemini 3.1 Pro	−7.3	−6.4	−8.7	−10.3	−5.6	−7.7
Gemini 3.1 Flash	−11.4	−10.2	−13.6	−15.1	−9.3	−11.9
Gemini 3 Pro	−9.6	−8.7	−11.4	−13.2	−7.8	−10.1
Gemini 3.1 Flash Lite	−14.8	−13.4	−16.9	−18.7	−12.6	−15.3
Llama 4	−12.6	−11.3	−14.8	−16.4	−10.7	−13.2
DeepSeek-V3	−14.1	−12.8	−16.3	−18.2	−12.1	−14.7
Gemma 3	−17.3	−15.8	−19.4	−21.6	−14.8	−17.8
Phi-4	−19.6	−17.9	−22.1	−24.3	−16.8	−20.1
Mistral Large V3	−18.4	−16.7	−20.8	−23.1	−15.6	−18.9
Falcon 3	−22.7	−20.4	−25.3	−27.8	−19.2	−23.1
DBRX Vision	−20.9	−18.6	−23.4	−25.7	−17.4	−21.2
GPT-4o (baseline)	−14.2	−12.8	−16.4	−18.6	−11.3	−14.7

Table 7. Multilingual Performance and Cross-Lingual Consistency (%). The Max Δ column indicates the maximum performance divergence across the evaluated languages (English, Spanish, Chinese, Hindi, Arabic), highlighting the cross-lingual robustness of each model.

Model	English	Spanish	Chinese	Hindi	Arabic	Max Δ
GPT-5.2	82.5	81.8	81.2	80.6	80.1	2.4
GPT-5.1	80.3	79.6	78.9	78.3	77.8	2.5
GPT-5	77.6	76.8	76.1	75.4	74.9	2.7
o3-pro	81.5	80.7	80.3	79.6	79.1	2.4
o3	78.7	77.9	77.4	76.8	76.2	2.5
Claude Opus 4.6	79.2	78.4	77.8	77.1	76.6	2.6
Claude Sonnet 4.6	75.6	74.8	74.1	73.4	72.9	2.7
Claude Opus 4.5	73.6	72.8	72.1	71.4	70.9	2.7
Claude 4.1	71.1	70.3	69.6	68.9	68.3	2.8
Gemini 3.1 Pro	81.5	80.8	81.1	79.9	79.3	2.2
Gemini 3.1 Flash	74.0	73.2	73.6	72.3	71.7	2.3
Gemini 3 Pro	78.0	77.2	77.6	76.4	75.8	2.2
Gemini 3.1 Flash Lite	67.1	66.2	66.7	65.3	64.6	2.5
Llama 4	72.0	70.8	69.6	68.4	67.3	4.7
DeepSeek-V3	69.7	68.2	70.1	67.3	65.8	4.3
Gemma 3	63.9	62.4	61.8	60.3	58.9	5.0
Phi-4	61.0	59.6	58.7	57.4	55.9	5.1
Mistral Large V3	62.8	61.3	60.4	58.9	57.6	5.2
Falcon 3	57.7	55.8	54.2	52.6	50.8	6.9
DBRX Vision	59.9	58.1	56.7	54.9	53.2	6.7
GPT-4o (baseline)	66.2	65.4	64.1	63.8	63.2	3.0

Table 8. Performance by Task Complexity (%). The Easy-Hard Gap highlights the degradation in model accuracy as reasoning requirements and visual complexity increase. Notably, reasoning-focused models (e.g. o3-pro) demonstrate a narrower performance gap.

Model	Easy	Medium	Hard	Easy–Hard Gap
GPT-5.2	93.6	83.4	64.8	28.8
GPT-5.1	91.4	81.2	62.4	29.0
GPT-5	88.7	78.6	59.8	28.9
o3-pro	92.3	82.7	67.4	24.9
o3	89.6	79.8	64.1	25.5
Claude Opus 4.6	90.2	80.3	61.7	28.5
Claude Sonnet 4.6	86.4	76.1	58.2	28.2
Claude Opus 4.5	84.1	73.8	56.4	27.7
Claude 4.1	81.6	71.3	54.1	27.5
Gemini 3.1 Pro	92.1	82.4	63.6	28.5
Gemini 3.1 Flash	84.6	74.2	56.8	27.8
Gemini 3 Pro	88.4	78.3	60.4	28.0
Gemini 3.1 Flash Lite	77.3	67.2	49.6	27.7
Llama 4	82.4	72.1	54.8	27.6
DeepSeek-V3	79.8	69.6	52.3	27.5
Gemma 3	73.6	63.8	46.9	26.7
Phi-4	70.4	61.2	44.3	26.1
Mistral Large V3	72.1	62.6	45.8	26.3
Falcon 3	66.8	57.4	40.2	26.6
DBRX Vision	69.3	59.8	42.7	26.6
GPT-4o (baseline)	78.4	65.8	48.3	30.1

Table 9. Aggregate Performance and Robustness by Model Family. This summary aggregates key metrics across frontier proprietary models and the open-source average by highlighting broad trends in overall accuracy, counterfactual robustness (Mean CAD) and cross-lingual consistency.

Metric	OpenAI (GPT-5.x + o3)	Anthropic (Claude 4.x)	Google (Gemini 3.x)	Open-Source	GPT-4o Baseline
Avg Accuracy (%)	80.1	74.9	75.1	64.5	66.2
Mean CAD (%)	−8.1	−11.6	−11.2	−19.0	−14.7
Best Bias Type	Text (89.9)	Text (88.2)	Text (89.3)	Text (74.9)	Text (78.2)
Worst Bias Type	Compound (67.1)	Compound (59.4)	Compound (58.9)	Compound (45.6)	Compound (48.6)
Cross-lingual Max Δ	2.5%	2.7%	2.3%	5.7%	3.0%
Easy Accuracy (%)	91.1	85.6	85.6	72.4	78.4
Hard Accuracy (%)	63.7	57.6	57.6	45.6	48.3
Easy–Hard Gap	27.4	28.0	28.0	26.8	30.1
Texture CAD	−10.7	−13.7	−11.8	−24.3	−18.6
Count CAD	−8.3	−11.2	−9.6	−21.7	−16.4

Table 10. Performance on Hard or Complex Bias Types and Reasoning Advantage (%). This table highlights the specific performance gains achieved by reasoning-augmented models (e.g. o3-pro) compared to standard unified multimodal transformers (e.g., GPT-5) on challenging cognitive bias categories.

Bias Type	**o3-pro**	**o3**	**GPT-5**	**Claude Opus 4.6**	**Gemini 3.1 Pro**	**Reasoning Advantage**
Texture	72.4	69.1	64.3	66.8	70.2	+8.1% (o3-pro vs GPT-5)
Counting	79.6	76.3	63.8	67.4	74.1	+15.8%
Spatial	71.8	68.4	62.7	65.3	69.6	+9.1%
Physical	68.3	65.2	59.8	62.4	66.7	+8.5%
Temporal	74.2	71.1	58.4	61.8	71.8	+15.8%
Spurious	70.6	67.4	63.1	66.2	68.4	+7.5%
Compositional	73.8	70.6	61.4	64.7	70.3	+12.4%
Typography	68.4	65.3	57.6	60.8	65.2	+10.8%
Compound	61.3	58.2	48.7	52.4	57.6	+12.6%
Avg (Hard)	**71.2**	**68.0**	**59.8**	**63.1**	**68.2**	+11.4%

Table 11. Performance Improvement (Δ) over Baseline by Bias Category. The table reports the absolute percentage-point increase in frontier models relative to the baseline across all cognitive bias types. The 'Best Model' column highlights the architectural advantage for specific reasoning tasks, notably showing the strength of reasoning-augmented models (e.g. o3-pro) on compositional and compound tasks.

Bias Type	**GPT-5.2 Δ**	**o3-pro Δ**	**Claude Opus 4.6 Δ**	**Gemini 3.1 Pro Δ**	**Llama 4 Δ**	**Best Model**
Texture	+14.2	+11.1	+10.6	+12.5	+4.4	GPT-5.2
Counting	+16.3	+18.4	+12.5	+14.6	+4.6	o3-pro
Spatial	+15.8	+12.6	+12.2	+14.3	+4.8	GPT-5.2
Physics	+17.3	+14.4	+13.8	+15.9	+6.3	GPT-5.2
Temporal	+17.1	+19.3	+14.5	+16.2	+7.1	o3-pro
Spurious	+14.2	+10.6	+10.8	+13.0	+3.6	GPT-5.2
Compositional	+17.0	+19.2	+14.2	+16.3	+6.7	o3-pro
Text	+13.1	+8.2	+10.0	+11.9	+3.2	GPT-5.2
Scale	+15.9	+12.5	+12.1	+14.5	+5.0	GPT-5.2
Cultural	+18.1	+15.5	+15.1	+17.3	+8.1	GPT-5.2
Occlusion	+16.5	+15.9	+13.6	+15.7	+6.6	GPT-5.2
Temp. Consist.	+17.2	+19.0	+14.4	+16.3	+7.2	o3-pro
Typography	+17.1	+13.4	+13.8	+16.0	+6.6	GPT-5.2
Compound	+18.7	+24.0	+15.1	+17.6	+7.0	o3-pro
Avg Δ	+16.2	+15.3	+13.0	+15.3	+5.8	—

Table 12. Worst Performing Bias Categories by Model. This table highlights the specific cognitive bottlenecks for each model. Across almost all architectures, compound reasoning (requiring simultaneous resolution of multiple bias types) and cultural understanding remain the most significant challenges with smaller open-source models falling to near-random performance on compound tasks.

Model	Worst Bias Type	Accuracy (%)	2nd Worst	Notes
GPT-5.2	Compound	67.3	Cultural (76.4)	Multi-step reasoning still hardest
GPT-5.1	Compound	64.8	Cultural (74.2)	—
GPT-5	Compound	61.4	Cultural (71.6)	—
o3-pro	Compound	72.6	Cultural (73.8)	Narrowest worst-bias gap of any model
o3	Cultural	70.9	Compound (69.3)	Cultural overtakes Compound
Claude Opus 4.6	Compound	63.7	Cultural (73.4)	—
Claude Sonnet 4.6	Compound	59.4	Cultural (69.8)	—
Claude Opus 4.5	Compound	57.6	Cultural (67.9)	—
Claude 4.1	Compound	54.8	Cultural (65.3)	—
Gemini 3.1 Pro	Compound	66.2	Cultural (75.6)	—
Gemini 3.1 Flash	Compound	57.1	Cultural (68.4)	—
Gemini 3 Pro	Compound	62.4	Cultural (72.8)	—
Gemini 3.1 Flash Lite	Compound	49.8	Cultural (61.3)	Near-chance on Compound
Llama 4	Compound	55.6	Cultural (66.4)	—
DeepSeek-V3	Compound	53.2	Cultural (64.1)	—
Gemma 3	Compound	46.7	Cultural (57.8)	Below chance threshold
Phi-4	Compound	43.8	Cultural (54.6)	Below chance threshold
Mistral Large V3	Compound	45.3	Cultural (56.4)	Below chance threshold
Falcon 3	Compound	40.2	Cultural (51.4)	Near-random on Compound
DBRX Vision	Compound	42.4	Cultural (53.6)	—

Table 13. Performance Degradation in Counting Tasks (Subitising Drop). Accuracy (%) is shown as a function of object count. The Subitising Drop" represents the performance degradation between the typical subitising limit (4 objects) and complex counting (8+ objects). Models with explicit reasoning pathways (e.g. o3-pro) exhibit significantly greater robustness as object counts scale past immediate visual apprehension limits.

Model	1 obj	2 obj	3 obj	4 obj	5 obj	6 obj	7 obj	8+ obj	Subitizing Drop
GPT-5.2	98.4	97.6	96.8	94.3	87.6	78.4	71.2	62.8	−31.5
o3-pro	98.7	98.1	97.4	96.2	92.4	86.3	80.1	73.6	−22.6
Claude Opus 4.6	97.8	96.9	95.7	93.1	84.8	74.6	67.3	58.9	−34.2
Gemini 3.1 Pro	98.2	97.4	96.3	93.8	86.4	76.8	69.4	61.2	−32.6
Llama 4	96.4	95.1	93.6	90.8	79.3	67.4	59.8	51.3	−39.5
DeepSeek-V3	95.8	94.2	92.4	89.3	76.8	64.2	56.7	48.4	−41.4
Gemma 3	93.4	91.6	89.3	85.7	70.4	57.8	49.6	41.3	−44.4
Phi-4	91.8	89.6	87.1	83.2	66.8	53.4	45.1	37.6	−46.7
Falcon 3	89.3	87.1	84.6	80.4	61.2	48.3	40.2	32.8	−48.5
GPT-4o (baseline)	95.3	93.8	91.6	88.4	74.2	62.1	54.3	46.7	−41.7

Table 14. Impact of Specific Counterfactual Perturbations on Bias Categories (CAD %). The table shows the isolated effect of targeted image edits on different reasoning tasks across averaged across all models. The strong diagonal trends (e.g, Object Removal heavily impacting Counting, Texture Strip impacting Texture) validate the precision of the counterfactual interventions.

Bias Type	**BG Swap**	**Color Shift**	**Obj. Removal**	**Texture Strip**	**Mirror Edit**
Texture	-4.2	-3.8	-2.1	-41.6	-5.3
Counting	-3.4	-2.9	-38.7	-8.4	-4.1
Spatial	-6.8	-4.3	-5.2	-9.6	-29.4
Physical	-7.3	-5.1	-4.8	-12.4	-18.6
Temporal	-8.6	-4.7	-6.3	-11.8	-21.3
Spurious	-32.4	-6.2	-8.1	-14.3	-7.6
Compositional	-8.4	-28.6	-9.2	-13.7	-6.8
Text-in-Image	-4.1	-3.6	-2.8	-7.4	-3.2
Scale	-5.6	-4.8	-3.7	-19.8	-8.4
Cultural	-21.3	-7.4	-6.8	-16.2	-9.1
Occlusion	-3.8	-2.9	-24.6	-18.3	-4.6
Typography	-6.2	-5.4	-4.1	-8.7	-3.9
Compound	-14.8	-12.3	-18.4	-26.7	-16.4

Table 15. Model Efficiency: Accuracy per Billion Parameters. This metric evaluates the architectural efficiency of open-weight models by normalising average accuracy against total parameter count (Acc/B Params). Smaller, dense models like Phi-4 achieve significantly higher information density than massive Mixture-of-Experts (MoE) architectures.

Model	**Parameters**	**Avg Accuracy (%)**	**Acc / B Params**	**Rank**
Phi-4	14B	61.0	4.36	1
Gemma 3	27B	63.9	2.37	2
Llama 4	405B	72.0	0.178	3
Mistral Large V3	123B	62.8	0.511	4
Falcon 3	40B	57.7	1.44	5
DBRX Vision	132B (MoE)	59.9	0.454	6
DeepSeek-V3	671B (MoE)	69.7	0.104	7

Table 16. Performance Gap between Proprietary and Open-Source Multimodal Models (%). The gap represents the absolute difference in accuracy across all bias categories. The largest disparities are observed in Counting and Compound reasoning and suggesting that while open-source models are closing the gap in standard perception (e.g. Texture), complex reasoning remains a significant frontier for proprietary research.

Bias Type	**Proprietary Avg (%)**	**Open-Source Avg (%)**	**Gap (%)**
Texture	85.3	71.1	14.2
Counting	83.3	64.1	19.2
Spatial	85.2	69.2	16.0
Physical	83.7	67.3	16.4
Temporal	80.9	63.4	17.5
Spurious	84.8	68.6	16.2
Compositional	82.1	64.5	17.6
Text-in-Image	89.0	73.2	15.8
Scale	83.6	67.7	15.9
Cultural	74.8	57.9	16.9
Occlusion	79.2	62.7	16.5
Temp. Consist.	78.7	61.8	16.9
Typography	77.6	60.6	17.0
Compound	67.5	48.9	18.6
Overall Avg	**81.7**	**64.5**	**17.2**

Table 17. Model Inter-Agreement and Error Correlation (%). Agreement Rate represents the frequency with which both models align the same answer (regardless of correctness). The "both Wrong" metric highlights the inherent difficulty of specific counterfactual samples where as failures are consistent across diverse architectures (e.g. 22.8% for Gemma 3 $\leftrightarrow$ Falcon 3).

Model Pair	**Agreement Rate (%)**	**Both Correct (%)**	**Both Wrong (%)**	**One Right / One Wrong (%)**
GPT-5.2 $\leftrightarrow$ o3-pro	89.4	81.6	7.8	10.6
GPT-5.2 $\leftrightarrow$ Claude Opus 4.6	84.7	76.3	8.4	15.3
GPT-5.2 $\leftrightarrow$ Gemini 3.1 Pro	86.2	78.4	7.8	13.8
Claude Opus 4.6 $\leftrightarrow$ Gemini 3.1 Pro	83.1	74.8	8.3	16.9
GPT-5.2 $\leftrightarrow$ Llama 4	76.4	63.7	12.7	23.6
GPT-5.2 $\leftrightarrow$ Falcon 3	68.2	54.8	13.4	31.8
Llama 4 $\leftrightarrow$ DeepSeek-V3	79.3	64.2	15.1	20.7
Llama 4 $\leftrightarrow$ Gemma 3	77.6	60.4	17.2	22.4
Gemma 3 $\leftrightarrow$ Phi-4	82.4	61.8	20.6	17.6
Gemma 3 $\leftrightarrow$ Falcon 3	80.1	57.3	22.8	19.9