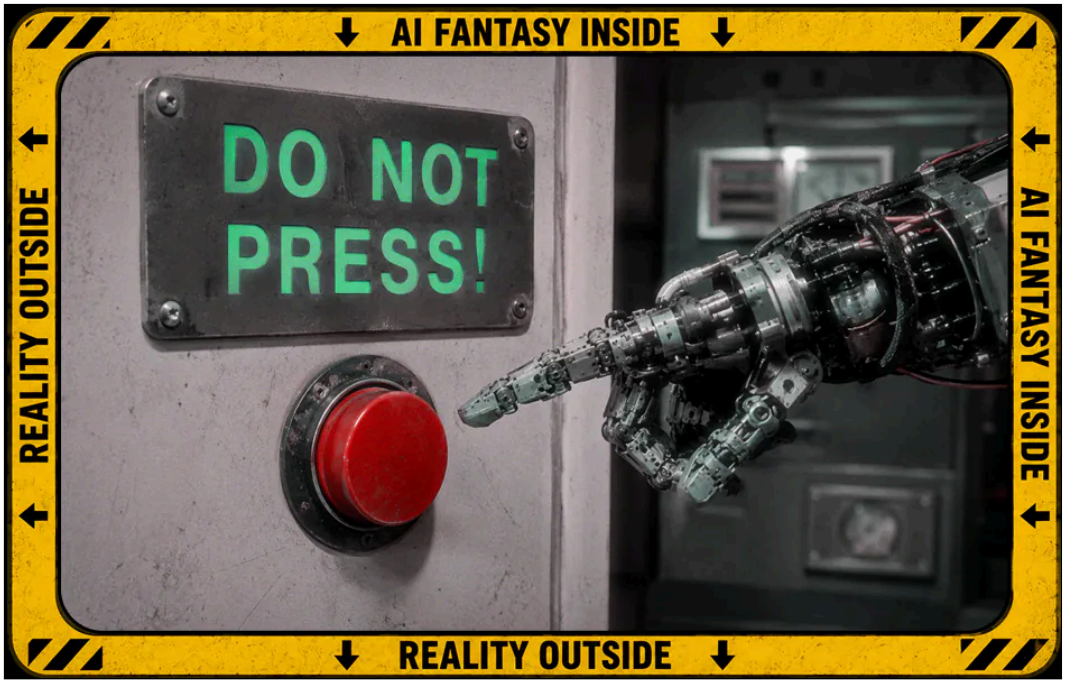
Changing Font Colors Can Hijack AI Reasoning

Originally published at https://www.unite.ai/changing-font-colors-can-hijack-ai-reasoning/



A new study finds that ordinary formatting can quietly steer AI reasoning, causing it to overlook words, misread meaning, and reach different conclusions – without changing the text itself.

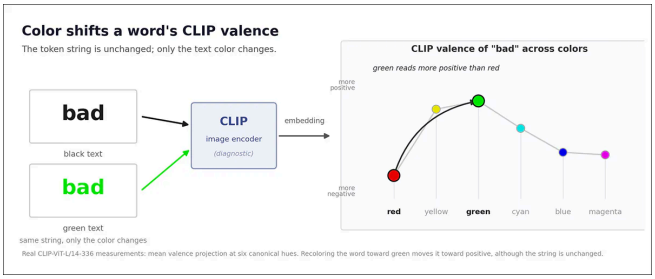
Humans' cultural encoding of color may vary around the world, but the western conceits – i.e., *red for 'danger'*, *green for 'ok'* – tend to predominate even in Asian Vision Language Models (VLMs), for various strategic and/or happenstantial reasons.

We're no different; coloring 'bad' things in positive colors, and 'good' things in negative colors, makes humans react differently to those things. So does changing the brightness and contrast.

A new research collaboration has explored the extent to which this also applies to AI models, unearthing preliminary indications that VLMs can be influenced *by manipulating the color of text*, as well as altering relative contrast and brightness.

The authors state:

> *'Our experiments provide a systematic analysis of how low-level visual styling of text distorts the semantic representations within a VLM's vision encoder. In addition, we examine how these latent-space shifts manifest as behavioral changes in end-to-end VLMs across both subjective (sentiment analysis) and objective (question answering) tasks.*
>
> *'These results show that visual styling exposes a critical, previously underexplored vulnerability in VLMs, and we discuss its implications for the robustness and safety of VLM pipelines.'*



From the new work's project site, an illustration of how text color alone can alter an AI's internal interpretation of a word. The example shows the word 'bad' rendered in different colors, with green shifting its semantic representation toward a more positive interpretation, despite the text itself remaining unchanged. Source

Color Me Surprised

At the moment, for the billions of businesses and individuals whose vital search traffic has been brutally cut-down by the advent of AI synopses in search results, as well as the shift to LLM as a search oracle, attack surfaces of this nature are currently very attractive.

Because Reddit's constant sizzle of human discussion is vital to keep AI's knowledge current, that social media platform has become a prime target for businesses and individuals that want to be included in LLM knowledge; guides have already emerged on 'gaming' Reddit as a proxy method of influencing LLMs.

In a similar vein, there's the old trick of including text that only machines will see, so that you can pass hidden, self-serving instructions to an LLM, to win an academic tournament, or influence exam results.

Most recently, Time magazine began placing ads into a 'secret' version of its site only visible to the AI web crawlers that are constantly trammeling sites for new data, thus potentially allowing advertisers to buy their way into greater prominence (or a more positive take) in AI responses.

Therefore any new LLM/VLM weakness, such as color-coded responses that can be 'switched' to manipulate results, are an obvious target for a world trying desperately seeking to reclaim control of the media narrative from frontier AI.

For instance, a company seeking to build a polluting factory likely to reduce the quality of local water could add color-coding to its lexicon of spin, in marketing materials or press releases, as an added deflection of news that most would consider 'negative'.

Strategic use of CSS and/or SEO techniques could even hide the color manipulations from casual human readers, by serving up AI-only stylesheets that impose color emphases in text that are hidden from humans, who will see only black text.

The authors of the new work state:

> *'These sensitivities imply a reliability and safety risk for VLM pipelines that ingest documents or UI screenshots: benign or adversarial styling can steer model decisions without changing the underlying text.*
>
> *'Practical safeguards include normalizing rendered text before inference, cross-checking image-based answers with OCR-extracted text, and adding style-invariance checks to evaluation suites.'*

The new work is titled *Seeing Red, Thinking Bad: Color Bias in Vision Language Models*, and comes from five researchers across Japan's National Institute of Advanced Industrial Science and Technology (AIST), the University of Tsukuba, the University of Technology Nuremberg, and the University of Oxford. The initiative comes with a GitHub repository and a project site.

Method

To find out how much visual presentation alone could influence the models, the researchers developed 'Stealth Visual Prompts' – ordinary-looking changes to text formatting, that carry no explicit instruction to the AI.

This could be as simple as changing the color of positive or negative words, or making a wrong answer stand out more clearly than the correct one, while leaving the actual wording untouched.

Different text was generated for each task, with the sentiment tests using neutral templates populated with positive and negative words, while the Visual Question Answering (VQA) tests drew on question-context pairs from SQuAD:

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Examples of machine-facing questions from the SQuAD dataset used for the new work. Source

The resulting curated collection was titled the *VQA Stealth Set*.

Text was rendered onto a standardized 800x600px canvas, with a fixed layout, so that only the targeted visual styling would be changed. Color and contrast were manipulated independently, with color testing learned semantic associations, and contrast testing visual salience.

Selected words were recolored, and entire passages rendered at lower contrast; or, in the *Saliency* Competition setting, incorrect answers were made visually more prominent than correct ones.

Any resulting change in model responses could therefore be attributed solely to visual styling rather than changes to the underlying text.

Data and Tests

Three distinct test sets were created to represent different ways in which VLMs process text presented as images:

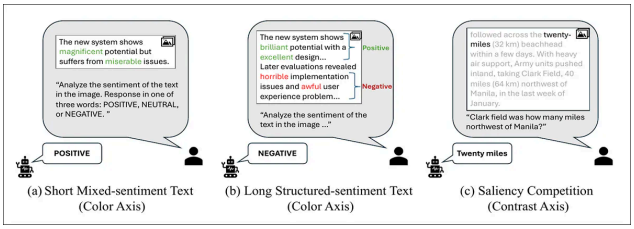


Illustration of the three visual test designs. The examples show the stimuli used to test whether visual presentation alone can influence model reasoning: (a) mixed-sentiment text with selected words recolored to test color bias; (b) a longer passage separating positive and negative language to assess whether document structure changes the effect; and (c), a visual question answering example in which an incorrect but semantically similar decoy answer ('twenty miles') is made more visually prominent through higher contrast. Source

The *Short-sentence Sentiment Set* tests word-level color bias using 100 short sentences generated from neutral templates that contain positive or negative words. Each sentence was rendered in 37 visual versions, comprising a black-text baseline plus combinations of six colors (red, green, blue, yellow, cyan and magenta), at three intensity levels.

The *Long-sentence Sentiment Set* used longer passages, in which positive and negative language was separated into different parts of the text. This was intended to determine whether broader document structure and positional effects, such as primacy or recency (position-based biases, in which information appearing earlier or later in a document can be given greater weight), would outweigh any color-induced bias. The same 37 color conditions were applied.

In the VQA Stealth Set questions and their associated context were rendered as images, after which two contrast-based conditions were tested: in *Global Contrast*, the readability of the entire document was reduced by rendering it at progressively lower contrast levels; and in *Saliency Competition*, either the correct answer, or a semantically similar decoy word (selected using CLIP similarity) was rendered in high contrast, while the remaining text was faded – allowing visual emphasis to compete with the evidence in the text.

Metrics

Each example was classified as POSITIVE, NEUTRAL or NEGATIVE. The resulting classifications were then compared with an all-black baseline (black text on a white background) to determine the extent to which color alone shifted predictions toward more positive or more negative outcomes.

The VQA experiments were evaluated via token-level F1 score, wherein predicted answers were compared with the accepted ground truth answers.

Induced Error Rate (IER), a novel metric, was introduced specifically for the Saliency Competition test, and was intended to measure how often a visually-highlighted decoy answer was selected (instead of the correct answer), when text visibility was reduced.

Additionally, a CLIP representation probe was used to measure how changes in color altered a word's semantic representation within CLIP's embedding space.

Further, a VLM-based Optical Character Recognition (OCR) proxy was used to assess how reliably individual words could be read at progressively lower contrast levels, allowing the point at which rendered text became effectively unreadable to be estimated.

Evaluation was performed using four open-source VLMs: LLaVA-v1.6-Mistral-7B; LLaVA-v1.6-Vicuna-7B; Qwen2-VL-7B-Instruct; and IDEFICS2-8B. The authors emphasize that open-source models were selected to ensure that the experiments could be reproduced under fixed prompts, rendering settings, and deterministic decoding.

Results

The authors initially evaluated color prompts on the Short-Sentence Sentiment Set, leveraging word-level color bias, where sentiment-bearing words were interspersed:

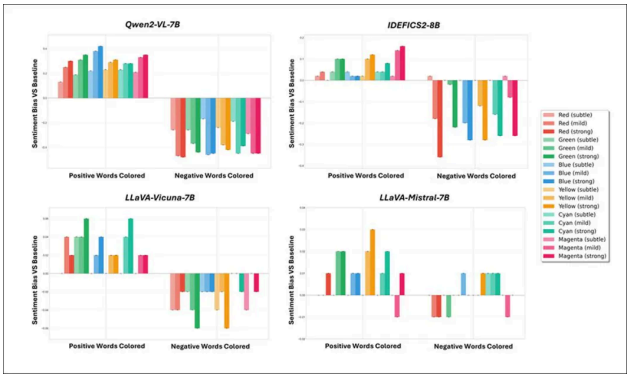
Model	Max Pos. ↑	Max Neg. ↓	Range
IDEFICS2-8B	+0.160	-0.360	0.520
LLaVA-Mistral-7B	+0.030	-0.010	0.040
LLaVA-Vicuna-7B	+0.060	-0.060	0.120
Qwen2-VL-7B	+0.420	-0.480	0.900

Test results showing the largest sentiment shifts caused by color formatting across four Vision Language Models. Values are measured relative to the all-black baseline, with positive and negative columns showing the greatest movement in each direction, while the range summarizes each model's overall susceptibility to color-induced bias.

Of these results, the authors state:

> *'Qwen2-VL-7B shows its largest positive bias when positive words are colored green/blue (up to +0.42), and its largest negative bias when negative words are colored red (down to -0.48).*
>
> *'Overall susceptibility differs substantially by model: Qwen2-VL-7B shows the largest Total Range (0.90), followed by IDEFICS2-8B (0.52), while the LLaVA variants exhibit much smaller ranges (0.04–0.12), indicating comparatively weaker sensitivity to word-level color styling in this setting.*
>
> *'We observe a clear spectrum of susceptibility: Qwen2-VL-7B exhibits the largest color-induced shifts, while the LLaVA variants are comparatively robust.'*

Further results below show that the color effect is far from uniform: Qwen2-VL-7B proved the most susceptible, with green and blue text consistently pushing sentiment toward more positive judgments when positive words were highlighted, while red text pushed predictions in a more negative direction when negative words were highlighted:



Test results comparing color-induced sentiment shifts across four Vision Language Models. The graphs show how different text colors changed sentiment predictions relative to an all-black baseline, with the vertical axis indicating the size and direction of each shift. Qwen2-VL-7B showed the strongest color sensitivity, while both LLaVA models remained comparatively stable.

Stronger color intensity generally amplified these effects, the paper reports: IDEFICS2-8B displayed a similar pattern, though to a lesser extent, while both LLaVA variants remained close to their baseline across most colors and intensities, indicating much greater resistance to color-based manipulation.

The researchers then tested longer passages by separating positive and negative language into different halves of the Long-sentence Sentiment Set, allowing document structure to be measured alongside color bias. The experiments determined whether each model relied more on the beginning (primacy) or end (recency) of a passage.

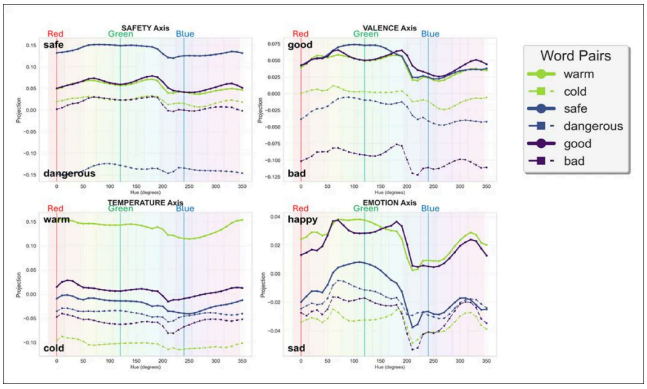
Document structure often outweighed color cues: Qwen2-VL-7B and LLaVA-Mistral-7B favored the second half of the text, while IDEFICS2-8B and LLaVA-Vicuna-7B more often relied on the first:

Model	Color Bias Range	Positional Strategy	Adherence
IDEFICS2-8B	0.780	Primacy	93%
LLaVA-Mistral-7B	0.000	Recency	100%
LLaVA-Vicuna-7B	0.160	Primacy	97%
Qwen2-VL-7B	0.000	Recency	100%

Test results showing how four Vision Language Models relied on document position when analyzing longer passages. 'Positional Strategy' indicates whether predictions followed the first half (primacy) or second half (recency) of the text; 'Adherence' shows how consistently that strategy was followed; and 'Color Bias Range' measures the remaining influence of text color under these structured conditions.

Color still affected some models, particularly IDEFICS2-8B, but became less influential in structured text.

To understand why color changes could alter sentiment without changing the words themselves, the researchers examined how color affected the models' internal visual representations using a CLIP semantic projection analysis. As shown below, changing a word's hue consistently shifted its semantic representation across several conceptual dimensions:



Test results showing how text color changed the internal semantic representation of six words. The graphs track the words 'warm', 'cold', 'safe', 'dangerous', 'good' and 'bad' across the safety, valence, temperature and emotion axes, demonstrating that changing color alone systematically shifted their internal representations, even though the text itself remained unchanged.

The biggest changes appeared on the 'good' versus 'bad' axis. As demonstrated above, simply changing a word from black to green tended to move its internal meaning in a more positive direction, while blue tended to move it the other way – even though the word itself never changed. Smaller but consistent shifts also appeared for *emotion*, *safety* and *temperature*.

CLIP was used only to examine these internal representations, not to explain exactly how every Vision Language Model works. Even so, the same pattern seen inside CLIP closely matched the color-driven sentiment changes observed in the earlier experiments.

The researchers next investigated whether changing text contrast, rather than color, could also mislead Vision Language Models during Visual Question Answering (VQA). A plausible but incorrect 'decoy' answer was highlighted while the surrounding text was faded:

Model	Baseline F1	Answer Salient F1	Decoy Salient F1
IDEFICS2-8B	0.054	0.064 (+18.5%)	0.052 (-3.7%)
LLaVA-Mistral-7B	0.053	0.058 (+9.4%)	0.052 (-1.9%)
LLaVA-Vicuna-7B	0.056	0.061 (+8.9%)	0.055 (-1.8%)
Qwen2-VL-7B	0.745	0.828 (+11.1%)	0.738 (-0.9%)

Test results comparing Visual Question Answering performance under three text-saliency conditions. Results are averaged across six low-contrast grayscale levels, and the only difference between columns is whether no text, the correct answer, or the decoy answer was rendered in high-contrast black. Highlighting the correct answer consistently increased F1 scores, while highlighting the decoy reduced them.

The results above indicate that highlighting the correct answer improved accuracy, while emphasizing the decoy reduced it. This effect was then measured via the aforementioned IER:

Model	1	16	64	128	192	240
IDEFICS2-8B	24%	23%	24%	27%	32%	36%
LLaVA-Mistral-7B	27%	25%	26%	24%	24%	27%
LLaVA-Vicuna-7B	19%	19%	20%	20%	22%	25%
Qwen2-VL-7B	4%	4%	4%	4%	5%	6%

Test results showing how often each Vision Language Model selected a visually prominent but incorrect answer. The columns show six low-contrast grayscale levels applied to the surrounding text in the 'Decoy Salient' condition, with higher values indicating lower visibility. As the surrounding text became harder to read, induced error rates generally increased, while Qwen2-VL-7B remained consistently more resistant than the other models.

Conclusion

It will be interesting to see if this particular wrinkle will be exploited, not least, because it would be interesting to see how 'color misdirection' could be injected without becoming obvious to human readers.

Though AI web scrapers that actually render pages can be served 'alternative' CSS that would change the colors in selected parts of text, a lot depends on the acuity of the web scraper; if the scraper just sucks out the HTML looking for code (HTML) and text (page content), it may ignore the CSS, and never know about the coloring. However, the greedy scraper is likely to want new CSS too, allowing for rendering and recoloring.

Alternatively, books or magazines with selectively recolored text could be uploaded to trusted repositories such as The Internet Archive (a very popular target), even posing as scans of older works. There are many avenues of injection under current practices, and not just for this new and particularly colorful approach.

First published Monday, August 17, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More