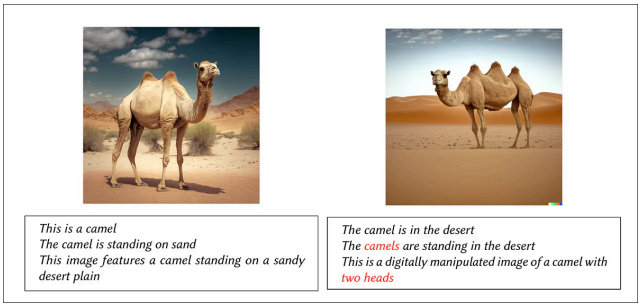
# Using AI Hallucinations to Evaluate Image Realism

Originally published at https://www.unite.ai/using-ai-hallucinations-to-evaluate-image-realism/

New research from Russia proposes an unconventional method to detect unrealistic AI-generated images – not by improving the accuracy of large vision-language models (LVLMs), but by intentionally leveraging their tendency to hallucinate.

The novel approach extracts multiple 'atomic facts' about an image using LVLMs, then applies natural language inference (NLI) to systematically measure contradictions among these statements – effectively turning the model's flaws into a diagnostic tool for detecting images that defy common-sense.




*Two images from the WHOOPS! dataset alongside automatically generated statements by the LVLM model. The left image is realistic, leading to consistent descriptions, while the unusual right image causes the model to hallucinate, producing contradictory or false statements.* Source: https://arxiv.org/pdf/2503.15948

Asked to assess the realism of the second image, the LVLM can see that *something* is amiss, since the depicted camel has three humps, which is unknown in nature.

However, the LVLM initially conflates *>2 humps* with *>2 animals*, since this is the only way you could ever see three humps in one 'camel picture'. It then proceeds to hallucinate something even more unlikely than three humps (i.e., 'two heads') and never details the very thing that appears to have triggered its suspicions – the improbable extra hump.

The researchers of the new work found that LVLM models can perform this kind of evaluation natively, and on a par with (or better than) models that have been fine-tuned for a task of this sort. Since fine-tuning is complicated, expensive and rather brittle in terms of downstream applicability, the discovery of a native use for one of the greatest roadblocks in the current AI revolution is a refreshing twist on the general trends in the literature.

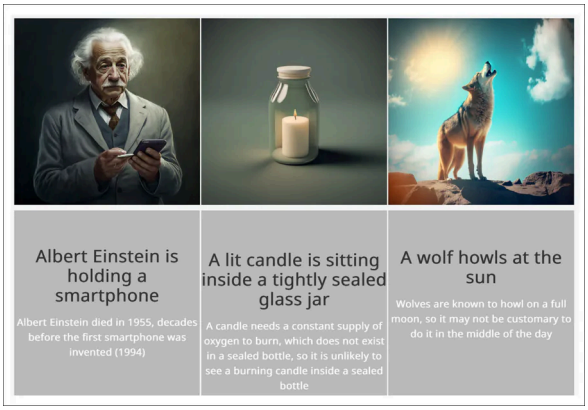
## Open Assessment

The importance of the approach, the authors assert, is that it can be deployed with *open source* frameworks. While an advanced and high-investment model such as ChatGPT can (the paper concedes) potentially offer better results in this task, the arguable real value of the literature for the majority of us (and especially for the hobbyist and VFX communities) is the possibility of incorporating and developing new breakthroughs in local implementations; conversely everything destined for a proprietary commercial API system is subject to withdrawal, arbitrary price rises, and censorship policies that are more likely to reflect a company's corporate concerns than the user's needs and responsibilities.

The new paper is titled *Don't Fight Hallucinations, Use Them: Estimating Image Realism using NLI over Atomic Facts*, and comes from five researchers across Skolkovo Institute of Science and Technology (Skoltech), Moscow Institute of Physics and Technology, and Russian companies MTS AI and AIRI. The work has an accompanying GitHub page.

## Method

The authors use the Israeli/US WHOOPS! Dataset for the project:

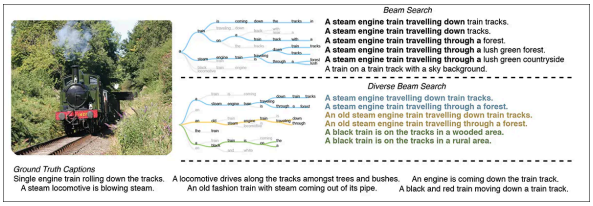



*Examples of impossible images from the WHOOPS! Dataset. It's notable how these images assemble plausible elements, and that their improbability must be calculated based on the concatenation of these incompatible facets.* Source: https://whoops-benchmark.github.io/

The dataset comprises 500 synthetic images and over 10,874 annotations, specifically designed to test AI models' commonsense reasoning and compositional understanding. It was created in collaboration with designers tasked with generating challenging images via text-to-image systems such as Midjourney and the DALL-E series – producing scenarios difficult or impossible to capture naturally:

*Further examples from the WHOOPS! dataset.* Source: https://huggingface.co/datasets/nlphuji/whoops

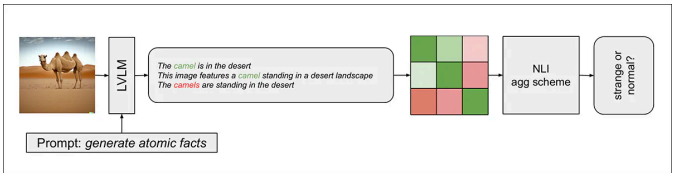
The new approach works in three stages: first, the LVLM (specifically LLaVA-v1.6-mistral-7b) is prompted to generate multiple simple statements – called 'atomic facts' – describing an image. These statements are generated using Diverse Beam Search, ensuring variability in the outputs.




*Diverse Beam Search produces a better variety of caption options by optimizing for a diversity-augmented objective.* Source: https://arxiv.org/pdf/1610.02424

Next, each generated statement is systematically compared to every other statement using a Natural Language Inference model, which assigns scores reflecting whether pairs of statements entail, contradict, or are neutral toward each other.

Contradictions indicate hallucinations or unrealistic elements within the image:



*Schema for the detection pipeline.*

Finally, the method aggregates these pairwise NLI scores into a single 'reality score' which quantifies the overall coherence of the generated statements.

The researchers explored different aggregation methods, with a clustering-based approach performing best. The authors applied the k-means clustering algorithm to separate individual NLI scores into two clusters, and the centroid of the lower-valued cluster was then chosen as the final metric.

Using two clusters directly aligns with the binary nature of the classification task, i.e., distinguishing realistic from unrealistic images. The logic is similar to simply picking the lowest score overall; however, clustering allows the metric to represent the average contradiction across multiple facts, rather than relying on a single outlier.

## Data and Tests

The researchers tested their system on the WHOOPS! baseline benchmark, using rotating test splits (i.e., cross-validation). Models tested were BLIP2 FlanT5-XL and BLIP2 FlanT5-XXL in splits, and BLIP2 FlanT5-XXL in zero-shot format (i.e., without additional training).

For an instruction-following baseline, the authors prompted the LVLMs with the phrase *'Is this unusual? Please explain briefly with a short sentence'*, which prior research found effective for spotting unrealistic images.

The models evaluated were LLaVA 1.6 Mistral 7B, LLaVA 1.6 Vicuna 13B, and two sizes (7/13 billion parameters) of InstructBLIP.

The testing procedure was centered on 102 pairs of realistic and unrealistic ('weird') images. Each pair was comprised of one normal image and one commonsense-defying counterpart.

Three human annotators labeled the images, reaching a consensus of 92%, indicating strong human agreement on what constituted 'weirdness'. The accuracy of the assessment methods was measured by their ability to correctly distinguish between realistic and unrealistic images.

The system was evaluated using three-fold cross-validation, randomly shuffling data with a fixed seed. The authors adjusted weights for entailment scores (statements that logically agree) and contradiction scores (statements that logically conflict) during training, while 'neutral' scores were fixed at zero. The final accuracy was computed as the average across all test splits.

| Model | # | Method | | |
|---|---|---|---|---|
| | | min | absmax | clust |
| nli-deberta-v3-large | 304M | 63.73 | 62.75 | **72.55** |
| nli-deberta-v3-base | 86M | 60.78 | 55.39 | **61.76** |
| nli-deberta-v3-small | 47M | **61.27** | 58.33 | 60.78 |

*Comparison of different NLI models and aggregation methods on a subset of five generated facts, measured by accuracy.*

Regarding the initial results shown above, the paper states:

> *'The ['clust'] method stands out as one of the best performing. This implies that the aggregation of all contradiction scores is crucial, rather than focusing only on extreme values. In addition, the largest NLI model (nli-deberta-v3-large) outperforms all others for all aggregation methods, suggesting that it captures the essence of the problem more effectively.'*

The authors found that the optimal weights consistently favored contradiction over entailment, indicating that contradictions were more informative for distinguishing unrealistic images. Their method outperformed all other zero-shot methods tested, closely approaching the performance of the fine-tuned BLIP2 model:

| Model | # | Mode | Acc ↑ |
|---|---|---|---|
| BLIP2 FlanT5-XL [4] | 3.94B | ft | 60.00 |
| BLIP2 FlanT5-XXL [4] | 12.4B | ft | **73.00** |
| BLIP2 FlanT5-XXL [4] | 12.4B | zs | 50.00 |
| LLaVA 1.6 Mistral 7B [13] | 7.57B | zs | 52.45 |
| LLaVA 1.6 Vicuna 13B [13] | 13.4B | zs | 56.37 |
| InstructBLIP [14] | 7B | zs | 61.27 |
| InstructBLIP [14] | 13B | zs | 62.25 |
| Ours (nli w/ clust agg) | 7.9B | zs | **72.55** |

*Performance of various approaches on the WHOOPS! benchmark. Fine-tuned (ft) methods appear at the top, while zero-shot (zs) methods are listed underneath. Model size indicates the number of parameters, and accuracy is used as the evaluation metric.*

They also noted, somewhat unexpectedly, that InstructBLIP performed better than comparable LLaVA models given the same prompt. While recognizing GPT-4o's superior accuracy, the paper emphasizes the authors' preference for demonstrating practical, open-source solutions, and, it seems, can reasonably claim novelty in explicitly exploiting hallucinations as a diagnostic tool.

## Conclusion

However, the authors acknowledge their project's debt to the 2024 FaithScore outing, a collaboration between the University of Texas at Dallas and Johns Hopkins University.

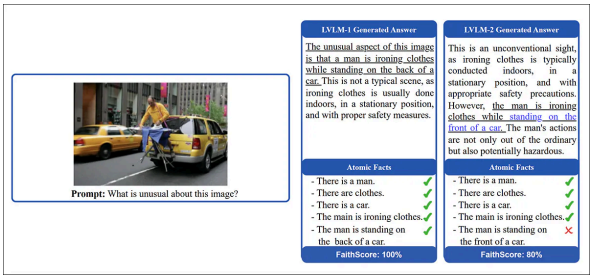



*Illustration of how FaithScore evaluation works. First, descriptive statements within an LVLM-generated answer are identified. Next, these statements are broken down into individual atomic facts. Finally, the atomic facts are compared against the input image to verify their accuracy. Underlined text highlights objective descriptive content, while blue text indicates hallucinated statements, allowing FaithScore to deliver an interpretable measure of factual correctness.* Source: https://arxiv.org/pdf/2311.01477

FaithScore measures faithfulness of LVLM-generated descriptions by verifying consistency against image content, while the new paper's methods explicitly exploit LVLM hallucinations to detect unrealistic images through contradictions in generated facts using Natural Language Inference.

The new work is, naturally, dependent upon the eccentricities of current language models, and on their disposition to hallucinate. If model development should ever bring forth an entirely non-hallucinating model, even the general principles of the new work would no longer be applicable. However, this remains a challenging prospect.

*First published Tuesday, March 25, 2025*

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**