The ‘Download More Labels!’ Illusion in AI Research

Originally published at https://www.unite.ai/the-download-more-labels-illusion-in-ai-research/

A common view in current machine learning research is that machine learning itself can be used to improve the quality of AI dataset annotations – particularly image captions intended for use in vision-language models (VLMs). This line of thinking is driven by the high cost of human annotation, and the added burden of supervising annotator performance.

Arguably this is the AI equivalent of the early 2000s ‘download more RAM’ meme, which satirized the notion that a hardware limitation could be resolved with a software-based fix.

It’s also an under-regarded issue; while new AI models attract widespread attention in both public and commercial spheres, annotation often appears to be a trivial detail in machine learning pipelines, overshadowed by the excitement surrounding broader frameworks.

In truth, the capacity of machine learning systems to recognize and reproduce patterns (the central use case of nearly all AI systems) is dependent on the quality and consistency of real-world annotations – labels and phrases which are created or adjudicated by real people, often making subjective judgments about individual data points in non-ideal circumstances.

Inevitably, systems which seek to observe and reproduce patterns in annotator behavior (and thereby replace human annotators and facilitate accurate labeling at scale) cannot hope to perform well on data *not* contained in the examples taken from human observers. Nothing ‘similar’ is quite the same, and cross-domain equivalency remains a problematic pursuit in computer vision.

The ‘upstream data buck’ has to stop somewhere, and in this case, that’s exactly where it stops – with a human cerebellum making some kind of subjective distinction in order to codify data for an artificial system.

The RAG Trade

Until recently, the inaccuracies arising from under-curated dataset annotations were, perhaps, seen as acceptable collateral damage in the context of the imperfect but still-marketable results obtained from generative AI systems.

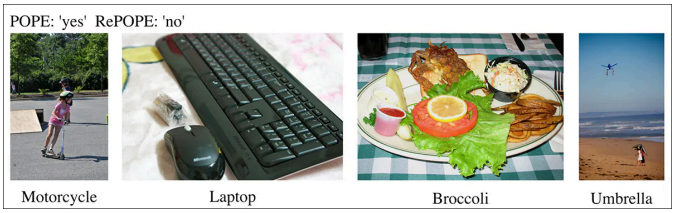
Indeed, only this year a study from Singapore concluded that hallucinations – i.e., the occasions when AI systems invent things that undermine our intentions – are inevitable, and bound in with the conceptual architecture of such systems.

To counter this, RAG-based agents – which can ‘verify’ facts through internet searches – are becoming popular in research and applied commercial solutions. However, they add to the resource cost and to the latency in queries; additionally, novel information applied to a trained model cannot compete with the more intricate and deeply-intertwined connections that characterize the native layers in a trained model.

It would therefore be better if the annotation data that informs these models was significantly less flawed in the first place, even if it cannot be perfect (not least because this activity encroaches into the realm of human subjectivity).

RePOPE

A new paper from Germany highlights the problems that arise from relying on older, widely used datasets, focusing in particular on the accuracy and reliability of their image captions. The researchers’ findings suggest that label errors in benchmarks can mask or misrepresent hallucination in vision-language models.



From the new paper, some examples where the original captions failed to correctly identify objects in the MSCOCO dataset of images. The researchers' manual revision of the POPE benchmark dataset addresses these shortcomings, demonstrating the cost of saving money on annotation curation. Source: https://arxiv.org/pdf/2504.15707

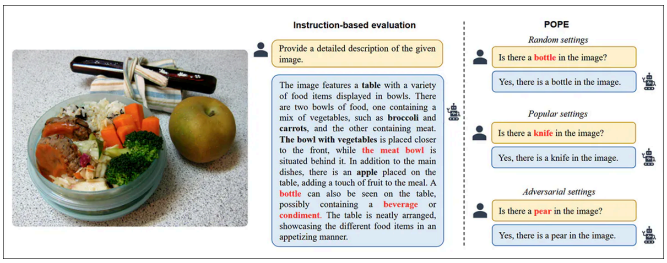
Imagine a model is shown an image of a street scene and asked whether there is a bicycle in it. The model answers *yes*. If the benchmark dataset says there is no bicycle, the model is marked *wrong*. But if a bicycle is *clearly visible* in the image, and was simply missed during annotation, then the model's answer was correct, and the benchmark has failed. Errors like this can accumulate across a dataset, giving a distorted picture of which models are accurate and which are prone to hallucination.

Thus, when incorrect or ambiguous annotations are treated as ground truth, models may appear to hallucinate when they are correct, or else seem accurate when they are not, distorting both the measurement of hallucination and the ranking of model performance, and making it harder to diagnose or address the problem with certainty.

The new paper revisits a widely used benchmark called Polling-based Object Probing Evaluation (POPE), which tests whether vision-language models can correctly say what is or isn't in an image.

POPE is based on labels from the influential Microsoft COCO: Common Objects in Context (MSCOCO) dataset, a collection of annotated images which has long been treated as offering a good level of annotation accuracy.

POPE evaluates object hallucination in large vision-language models by reframing the problem as a binary classification task. Rather than parsing generated captions, the system poses simple *yes/no* questions to the model about whether specific objects are present in an image, using templates such as *'Is there a <object> in the image?'*.



Examples of object hallucination in vision-language models. Bold labels indicate objects marked as present in the original annotations, while red labels show objects hallucinated by the models. The left example reflects a traditional instruction-based evaluation, while the three examples on the right are drawn from different POPE benchmark variants. Source: https://aclanthology.org/2023.emnlp-main.20.pdf

Ground-truth objects (answer: *Yes*) are paired with sampled non-existent objects (answer: *No*), chosen through random, frequent (*popular*), or co-occurrence-based (*adversarial*) strategies. This setup allows for more stable, prompt-insensitive evaluation of hallucination without relying on complex rule-based caption analysis.

The authors of the new paper – titled *RePOPE: Impact of Annotation Errors on the POPE Benchmark* – challenge the assumed accuracy of POPE by rechecking the labels on the benchmark's images (i.e., MSCOCO) – and finding that a surprising number are wrong or unclear.

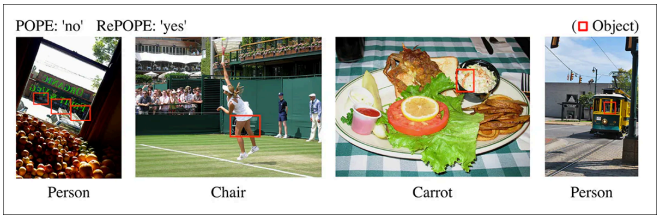
Examples from the 2014 MSCOCO dataset. Source: https://arxiv.org/pdf/1405.0312

These errors change the way models are ranked, with some that initially performed well falling behind when judged against corrected labels.

In tests, the authors evaluated a range of open-weight vision-language models on both the original POPE benchmark and their re-labeled *RePOPE* version.

According to the paper, the corrected annotations led to notable changes in model rankings, particularly in F1 scores, with several high-performing models under POPE dropping in position under RePOPE.

The authors contend that this shift illustrates the extent to which annotation errors can obscure the actual hallucination behavior of models, and they present RePOPE as a more reliable tool for assessing hallucination vulnerability.



In another example from the new paper, we see how the original POPE captions fail to discern subtle objects, such as a person sitting beside the cabin of a tram in the rightmost photo, or the chair obscured by the tennis player in the second photo from the left.

Method and Tests

The researchers re-labeled all the annotations in the original MSCOCO dataset, with two human labelers assigned to each data instance. Where ambiguity as to the quality of the original labels arose (as in the examples below), these results were set aside from the testing round.

Ambiguous cases, where labeling inconsistencies in POPE reflect unclear category boundaries. For instance, a teddy bear labeled as a bear, a motorcycle as a bicycle, or airport vehicles as cars. These cases were excluded from RePOPE due to the subjective nature of such classifications, as well as the inconsistencies in MSCOCO's original labels.

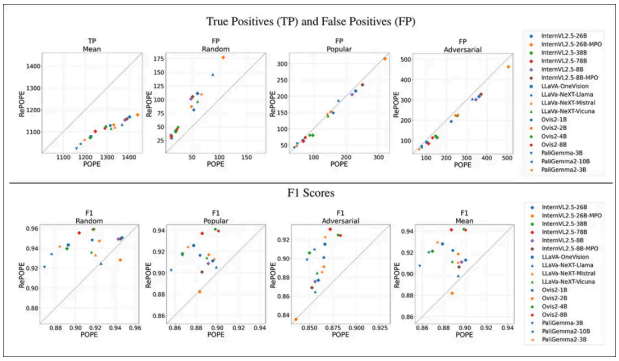
The paper states:

> *'The original annotators missed persons in the background or behind glass, the tennis player occludes the 'chairs' in the background and the cole slaw contains only a small visible stripe of a carrot.*
>
> *'For some objects, the COCO annotations are highly inconsistent likely due to differing definitions of those objects used by the original annotators. The classification of a 'teddy bear' as a 'bear', a motorcycle as a motorized 'bicycle', or an airport vehicle as a 'car' depends on specific definitions, leading to inconsistencies in POPE ground truth annotations. Therefore, we annotate the corresponding image-question pairs as 'ambiguous'.'*

	POPE: Yes			POPE: No		
RePOPE Labels	Yes	No	Amb.	Yes	No	Amb.
Random				0.3%	98.4%	1.3%
Popular	76.9%	9.3%	13.8%	2.6%	93.0%	4.4%
Adversarial				2.2%	90.5%	7.3%
Total	76.9%	9.3%	13.8%	1.7%	94.0%	4.3%

Results of the re-annotation: the positive questions are shared across all three POPE variants. Among those labeled 'Yes' in POPE, 9.3 percent were found to be incorrect and 13.8 percent were classified as ambiguous. For the 'No' questions, 1.7 percent were mislabeled and 4.3 percent were ambiguous.

The authors evaluated a range of open-weight models on POPE and on RePOPE, across diverse architectures and model sizes. The models chosen included some of the leading architectures on the OpenVLM leaderboard: InternVL2.5 (8B/26B/38B/78B and 8B-MPO/26B-MPO); LLaVA-NeXT; Vicuna; Mistral 7b; Llama; LLaVA-OneVision; Ovis2 (1B/2B/4B/8B); PaliGemma-3B; and PaliGemma2 (3B/10B).



Initial results: the high error rate in the original positive labels leads to a sharp drop in true positives across all models. False positives vary across subsets, nearly doubling on the random subset, but remaining largely unchanged on the popular subset, and show a slight decrease on the adversarial subset. The relabeling has a major effect on F1-based rankings. Models like Ovis2-4B and Ovis2-8B, which performed well on the popular and adversarial splits in POPE, also rise to the top on the random subset under RePOPE.. Please refer to the source PDF for better resolution.

The results graphs above illustrate how the number of true positives and false positives changes after correcting the labels in the benchmark.

True positives fell across all models, showing that they were often credited for correct answers when those answers were only correct under faulty labels, while false positives followed a more varied pattern.

On the 'random' version of POPE, false positives nearly *doubled* for many models, indicating that a significant number of objects flagged as hallucinations were actually present in the images but had been missed in the original annotations. In this case, many supposed model errors were in fact dataset labeling mistakes.

For the 'adversarial' version of POPE, where questions were based on objects that frequently co-occur, false positives decreased. This likely reflects a higher chance that the supposedly absent object was *actually in the image* but left *unlabeled*.

Although these shifts affected precision and recall, model rankings stayed relatively stable for both metrics.

The F1 score – POPE's main evaluation measure – was far more sensitive to the label corrections. On the random subset, models that ranked near the top under the original labels, such as InternVL2.5-8B and -26B, dropped to the bottom when scored with RePOPE. Others, such as Ovis2-4B and -8B, rose to the top.

A similar pattern emerged in the accuracy scores, though the authors note that these may now be biased, as the corrected dataset contains an uneven number of positive and negative examples.

The authors argue that the strong impact of annotation errors on benchmark results underscores the need for high-quality data. To support more reliable evaluation of object hallucination, they have released the corrected labels at GitHub.

However, they note that this re-labeling does not fully address the benchmark's saturation, since many models still achieve true positive and true negative rates above 90%. They suggest that additional benchmarks, such as DASH-B, which uses a more challenging set of negative examples, should be used alongside RePOPE.

Conclusion

This particular experiment was possible because of the very small scale of the dataset involved. Proving the same hypothesis on hyperscale datasets would involve working on very limited fragments of the data; in highly diverse large datasets, it might prove near-impossible to isolate statistically representative and semantically coherent groupings – potentially skewing the results.

Even if it were possible, what remedy would there be under the current state-of-the-art? The argument moves back inevitably towards the need for better and more copious human annotation.

In this regard, 'better' and 'more copious' exist as separate problems in their own right, since one can obtain a greater volume of annotations through race-to-the-bottom economies such as Amazon Mechanical Turk (AMT). Obviously, this potentially exploitative sub-economy frequently leads to inferior results.

Alternatively, one could farm out annotation tasks to economic regions where the same expenditure would yield a larger quantity of annotations. However, the further removed the annotator is from the intended use case of the model their labels will shape, the less likely it is that the resulting model will align with the needs or expectations of the target domain.

This therefore remains one of the most persistent and unresolved challenges in the economics of machine learning development.

First published Wednesday, April 23, 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