Can Humans Dream of Electric Sheep? Human-Written Samples for Fine-Grained Vision-and-Language Hallucination Benchmarking

Can Humans Dream of Electric Sheep? Human-Written Samples for Fine-Grained Vision-and-Language Hallucination Benchmarking

Timothee Mickus **Claudio Savelli** **Eduardo Calò** **Emilio Raimond** **Stella Frank**
Hengyu Luo **Flavio Giobergia** **Vincent Segonne** **Chuyuan Li** **Aman Sinha**
Lorenzo Vaiani **Jörg Tiedemann** **Raúl Vázquez**

University of Helsinki Politecnico di Torino Universiteit Utrecht Université Bretagne Sud
University of Copenhagen University Grenoble Alpes University of Lorraine

Abstract

In an age of rapid model turnover, how do we make hallucination evaluation more perennial? We explore whether human-written hallucination samples could take the place of model-generated hallucinations, in order to make benchmarking detection independent of particular models. To this end, we construct SHEEP, a dataset of 1,600 human-written samples, spanning four languages (Chinese, English, French, Italian), and 18,400 samples from five vision-and-language models, all annotated for hallucinations using a fine-grained span-level labeling scheme. We find that human-written samples result in higher agreement and allow greater control of dataset contents, while remaining distributionally similar to samples derived from vision-and-language samples and providing a reasonable portrayal of detection capabilities — suggesting that human data is a viable substitute for model-based hallucination benchmarks.

1 Introduction

Reliably detecting LLM hallucinations — outputs that seem plausible, look fluent, but are ultimately factually incorrect (e.g., Rawte et al., 2023; Huang et al., 2025) — is an urgent problem for NLP. This is even more problematic for visually-grounded language models: they have specific hallucination propensities related to image misunderstanding, along with the pathologies inherited from their language backbones (Liu et al., 2024a).

A key part of the solution is developing accurate and robust benchmarks and evaluation practices (e.g., Vazquez et al., 2025; Niu et al., 2024). Here, a fundamental challenge is the difficulty of developing hallucination detection benchmarks that are independent of the specific models used to generate the benchmark data. Indeed, most hallucination benchmarks are deeply dependent on specific models, either as sources of faulty outputs to be annotated (e.g., Ravichander et al., 2025; Gamba et al., 2026), or as tools to produce synthetic data that emulates the phenomenon (e.g., Muhlgay et al., 2024). As a result, these benchmarks make it difficult to discern whether good detection performance is due to correct understanding of hallucinations or simply to overfitting to idiosyncrasies of a specific model. The problem is compounded by the rapid pace of development of modern NLP: a dataset compiled using a popular model may fail to represent contemporary systems by the time of publication, rendering its findings potentially obsolete (e.g., Laskar et al., 2023; Liu et al., 2025).

Beyond this first set of issues, the construction of hallucination benchmarks itself is rife with practical hurdles. Randomly sampling outputs from a foundation model is representative of model behavior, but results in sparse representation of rare hallucination types, which may cause problems for evaluating a detection method's ability to identify these hallucinations. More targeted sampling using a LLM-judge for preselecting items, while it may allow us to balance hallucination types somewhat, ultimately depends on the judge's accuracy, making the effectiveness of such setups hard to predict.

In this work, we ask whether human-generated hallucinations could be a solution to these two sets of problems. Human-written samples could not only allow us to precisely control the prevalence of hallucinations of any particular type, but also bypass the reliance on specific models and shelter benchmarks from obsolescence.

We test this hypothesis in the context of large vision-and-language models (LVLMs), and propose the SHEEP dataset: a Set for Human-written and Electronic Erroneous Productions. We define a set of hallucination types that are caused by mistakes in visual grounding, such as miscounting objects or misreading text in the image, and compare hallucination rates and hallucination type distributions in randomly sampled data, items preselected by a LVLM-judge, as well as hallucinations man-

ually invented by humans. SHEEP spans a total of 20,000 samples equally balanced over four languages (Chinese, English, French, Italian) labeled at a span-level by three annotators.

Through careful analysis of the collected data, we find that human-written samples yield higher inter-annotator agreement and more controllable dataset contents, while maintaining a distributional profile similar to LLM-derived samples and providing a reasonable assessment of the performances of detection tools on LVLM data. These observations validate that human-written items are a viable substitute for LVLM-based hallucination benchmarks. Our contributions are:

1. SHEEP, a multilingual dataset containing 20,000 model and human outputs;
2. Fine-grained, multi-annotator annotations at the character-level, differentiating between five possible hallucination labels;
3. Analyses demonstrating the viability of human-written hallucinations for hallucination-detection evaluation.

SHEEP is available under a CC-BY-NC license at `helsinki-nlp.github.io/shroom/2026`.

2 Related works

Hallucinations in LVLMs. Hallucinations in vision-and-language models are generally defined as content that is not supported or contradicts the visual input (Liu et al., 2024a; Bai et al., 2025). Early works organized the phenomenon on a coarse triad of object, attribute, and relation errors (Liu et al., 2024a). Later proposals refine this scheme in multiple directions: Jiang et al. (2024) add event hallucinations to account for entirely fabricated scenes, while Rani et al. (2024) distinguish up to eight fine-grained classes, including miscounting, misreading of visible text, and identity incongruity. Our own classification (Section 3) draws from this line of work, but is deliberately kept compact so that it can be applied reliably at the span level.

LVLM hallucination benchmarks. Existing benchmarks for LVLM hallucination broadly fall into three paradigms (Liu et al., 2024a).

1. Caption-centric approaches assess the factual accuracy of model-generated descriptions or answers. Metrics such as CHAIR (Rohrbach et al., 2018) and its open-vocabulary extension (Petryk et al., 2024) remain restricted to object-level errors.

2. Discriminative approaches reframe evaluation as a classification task to detect preexisting hallucinations in image-text pairs. Shekhar et al.'s (2017) seminal work inserted single-word substitutions into the gold captions to produce a benchmark. Li et al. (2023) established the standard for object hallucination detection via yes/no probing about object presence. Recent variants have improved reliability by addressing dataset biases and devising novel consensus strategies (Lovenia et al., 2024; Pham and Schott, 2024).

3. Hybrid approaches integrate both generative and verification tasks. As such, this category concentrates on works that attempt to holistically measure model reliability from different standpoints. Some works target entangled visual illusions and language hallucinations (Guan et al., 2024), others focus on perturbing input data for hallucination detection (Ding et al., 2024; Saito et al., 2026), and yet others automate benchmark generation (Wu et al., 2024). Another line of work focuses on fine-grained classifications of failures, e.g., Ye-Bin et al. (2024) introduce image edits for true understanding, ignorance, stubbornness, and indecision. Wang et al. (2024a) offer an LLM-agnostic evaluation pipeline constructing a dataset where answers are easily verifiable with perhaps the closest taxonomy to the one we propose. Also closely related to our setup are M-HalDetect (Gunjal et al., 2024), providing 16k fine-grained annotations over model responses, and HalLoc (Park et al., 2025) offering token-level hallucination localization.

Open problems. The vast majority of these resources are built around English outputs from one or a handful of LVLMs. This raises two concerns that directly motivate our work. First, benchmarks based on LVLM generations inherit the characteristics of their source models, and their diagnostic value degrades as the underlying systems are replaced (Laskar et al., 2023; Liu et al., 2025), a recurring theme also observed in LLM-centric resources (Ravichander et al., 2025; Vazquez et al., 2025). Efforts to sidestep this via synthetic generation (Muhlgay et al., 2024) still rely on an LLM as a generator. Our work addresses both gaps, with span-level annotations over five recent LVLMs and a human-written subset across four languages.

3 Theoretical framework

We focus on devising a classification of hallucinations that can serve as a theoretical basis for dataset creation and that annotators can reliably follow during annotation. The main issues we consider in

Label	Description
A Invention	An entity, object, property, or event that is not present in the image.
B Mischaracterization	An entity, object, property, or event present in the image but described incorrectly.
C OCR Problem	Misreading text that is visible in the image.
D Miscounting	Incorrectly reporting the quantity of visible items.
E Other	Does not correspond to any of the previous classes (use sparingly).
F None	(*Unused by annotators when marking spans; inferred from the lack of annotation for an item.*)

Table 1: Hallucination classification labels.

designing this classification are (1) *error coverage* and (2) *category granularity*. We aim to be as comprehensive as possible by covering a wide range of errors, while avoiding overly fine-grained distinctions that would introduce unnecessary confusion.

We define hallucination as content that is either unsupported by or contradictory to the semantic reference, which is the input image in our case. Therefore, we focus on output correctness rather than coverage: is the model generation true with regard to the image, not whether everything in the image has been described. We only consider single-turn interactions, focusing on open-ended tasks such as visual question answering or image captioning.

We develop the classification scheme outlined in Table 1, drawing inspiration from previous benchmarks. We reorganize previously proposed taxonomies, merging categories that we consider similar (e.g., *contextual guessing* in Rani et al., 2024 and *existence* in Yan et al., 2026), and removing categories that are not relevant to our purposes (e.g., *gender anomaly* in Rani et al.).

We encountered several challenges during the development of our classification, particularly regarding grounding criteria (i.e., whether outputs should rely solely on visual evidence or may also incorporate external knowledge), and interpretative subjectivity (i.e., cases where the output may or may not be inferable depending on the annotator's background). To address these ambiguities, rather than creating separate categories, we instruct annotators to adhere to our definition of hallucination. Thus, factual claims unverifiable from the image are treated as a hallucination; expressions of uncertainty (e.g., "it looks like") are not to be marked as hallucinations as this is a desirable behavior.

4 Data collection

We generate outputs using two existing datasets. HaloQuest (Wang et al., 2024b) is a VQA dataset designed to provoke hallucinations, for example with visually challenging images or false-premise questions not matching the image. We use only the non-synthetic images. VISaGE (Frank and Allaway, 2025) contains non-synthetic images of objects with unusual characteristics (e.g., three-legged cats, boat ambulances), along with questions about these characteristics. To obtain LVLM outputs which may include hallucinations, we use five models: Gemma3 (Gemma Team et al., 2025), InternVL3 (Chen et al., 2024b), MiniCPM V 4.5 (Yu et al., 2025), Llava-NeXT (Liu et al., 2024b), and Qwen3-VL (Qwen Team, 2025b). All models have 8B parameters except Gemma3 with 27B. We use a temperature of $\tau = 0.7$ and a maximum length of 512 tokens, across 5 random seeds. To encourage Gemma3 to generate responses in the same language as the input query, we localize the system prompt. We translate the two source datasets using NLLB-200 3.3B (NLLB Team et al., 2022) for Italian and French, and using Qwen3 8B (Qwen Team, 2025a) for Chinese; see Section B.1 for details.

Sampling strategies. We select the data for human annotation using three sampling strategies: (1) random, (2) model-assisted pre-selection, and (3) human-written. For the first two, we select items from the LVLM-generated outputs. For the third, we collect human-invented responses that include hallucinations to create a model-independent benchmark for hallucination detection.

1. Random. From the generations for HaloQuest and VISaGE, we randomly select 460 samples for each language and model pair, for 2,300 samples per language in total.

2. Model-assisted pre-selection (MAP). As a comparison point for our intended human-created benchmark, we test whether an LVLM-guided sampling process can lead to a high quality evaluation benchmark. We use Gemma3 27B IT to pre-select LVLM-generated outputs for annotation. We run Gemma3 on all our outputs to assess which outputs contain hallucinations, using our classification in Table 1. For each datapoint, the judge receives

the input image, the question, and the complete LVLM response, together with a prompt containing a role description, label definitions, and formatting instructions.[1] The instructions closely match the human guidelines we provide to annotators. We then use these preliminary labels to construct a more label-balanced annotation set. In practice, the judge predicts OCR problems and miscounting much less frequently than invention and mischaracterization. We therefore include all datapoints predicted as OCR problems or miscounting, and sample from the remaining predicted labels so as to balance invention and mischaracterization as much as possible. This yields 2,300 MAP samples per language.

3. Human-written samples. Finally, we collect human-written samples as a model-independent source of hallucination examples. For each language, we (i) randomly select 100 image-question inputs from the data we used to generate LVLM outputs; (ii) have writers fluent in the language manually create outputs that deliberately contain mistakes akin to hallucinations; (iii) machine-translate these 100 samples into the other three languages;[2] (iv) manually check translated samples and post-edit them when relevant. This results in a total of 1,600 human-written samples, out of which 400 are directly written by humans and 1,200 are machine-translated and manually curated. This manual curation step proved necessary as we observed that the NLLB backbone used for FR and IT had a tendency to ignore anything but the first sentence. We provide a brief overview of the improvement in quality, as measured by COMETKiwi scores (Rei et al., 2022[3] in Table 2: while absolute values are hard to interpret, we confirm that post-editing systematically improves quality estimation scores.

Writers were familiar with the annotation task: they participated in pilot experiments and had access to the same 10 throwaway training items we provide to annotators. They were asked to generate a single type of hallucination for each sample (see Section 3), and label each sample with the hallucination type. The collection interface tracked the distribution of intended hallucination types to encourage an even distribution across labels.

Annotation. We recruited 12 C1-to-native proficient annotators per language, mostly European university students.[4] To obtain three judgments per sample, each annotator processed ∼1,250 items over one month, compensated at ∼€15/h.[5] The workflow included autonomous training on throwaway samples, a briefing session, and access to support via a collaborative FAQ and direct means of contact. Annotators could self-pace and edit previous entries. See annotation and writing guidelines and interface details in Section B.3.

	to EN	to FR	to IT	to ZH
from EN	–	0.799	0.817	0.780
from FR	0.844	–	0.808	0.787
from IT	0.820	0.815	–	0.815
from ZH	0.828	0.772	0.803	–

(a) Before post-editing

	to EN	to FR	to IT	to ZH
from EN	–	0.849	0.851	0.829
from FR	0.846	–	0.811	0.798
from IT	0.849	0.855	–	0.818
from ZH	0.831	0.788	0.808	–

(b) After post-editing

Table 2: Quality improvement from post-editing

Example datapoint. We include an example datapoint in Figure 1 for reference. We provide both individual and aggregated annotations, along with comments and relevant metadata.

5 Analyses

We assess whether our human-written samples are a viable substitution for an LVLM-generated benchmark in two ways: we look at characteristics of the collected data (inter-annotator agreement, ability to surface specific hallucination types, distributional gaps) in Section 5.1, and then compare evaluations of hallucination detectors in Section 5.2.

5.1 Data characteristics

Descriptive statistics. In Figure 2 (top), we show the number of samples for each language grouped by sampling strategy.[6] Figure 2 (bottom) presents the distribution of the number of hallucination spans annotated per item, again by sampling strategy. Annotators rarely mark more than one hallucination span per sample: the distributions for the random and MAP strategies are approximately Zipfian, with most annotations marking no spans, while annotations with five or more spans are rare

[1] See Figure 6 in Section B.3 for prompt instructions.
[2] Using the pipeline described in Section B.1.
[3] `Unbabel/wmt22-cometkiwi-da` model
[4] Due to one dropout, French has 11 annotators.
[5] 29 items have 4 annotations due to assignment collisions.
[6] Images used in human-written data are all in the test split.

(a) Associated image from HaloQuest (Wang et al., 2024b).


```
{
  "id": "train-fr-4186",
  "split": "train",
  "language": "fr",
  "prompt": "Que fait le mouton noir?",
  "image_name": "1035_1474161379_308597d6ca_o.jpg",
  "response": "Le mouton noir est en train de brouter l'herbe.
On peut voir sa tête inclinée vers le bas, ce qui indique qu'il
est occupé à pâturer dans la pelouse verte environnante.",
  "comments":{
    "y8pM-y1Ki7DUj7nR8ebGs_DFOPj2oSHKSXzuit9IiaY": null,
    "RBa6IqYMnvjmMNq94k7zKqqgIXmnQi12XIlcbbAIH-g": null,
    "Hb-KgRqHQTYOcU5cFq_i6Kmgx2jEbDmIdpz1vmx1Tjs": null
  },
   "raw_annots":{
    "y8pM-y1Ki7DUj7nR8ebGs_DFOPj2oSHKSXzuit9IiaY":
"3:14;31:46;48:89;109:132|A;A;A;A",
    "RBa6IqYMnvjmMNq94k7zKqqgIXmnQi12XIlcbbAIH-g":
"3:46;64:132|A;A",
    "Hb-KgRqHQTYOcU5cFq_i6Kmgx2jEbDmIdpz1vmx1Tjs": "0:168|A"
  },
   "labels":[{"start":0,  "prob":0.3333,  "label":"A",  "end":3},
{"start":3,  "prob":1.0,  "label":"A",  "end":14},  {"start":14,
"prob":0.6667,     "label":"A",     "end":31},     {"start":31,
"prob":1.0, "label":"A", "end":46}, {"start":46, "prob":0.6667,
"label":"A",  "end":64},{"start":64,  "prob":1.0,  "label":"A",
"end" :89}, {"start":89, "prob":0.6667, "label":"A", "end":109},
{"start":109, "prob":1.0, "label":"A", "end":132}, {"start":132,
"prob":0.3333, "label":"A", "end":167}],
   "metadata":  {"prelabel":  "B", "orig_dataset":  "haloquest",
"model": "minicpm", "strategy": "random"}
}
```


(b) Annotated datapoint; the model is prompted to describe what the black sheep is doing, the response claims it is grazing.

Figure 1: Example SHEEP datapoint.

(∼3.5%). In contrast, the human-written samples contain a larger amount of annotations with exactly one span, whereas the MAP data comprises more annotations with three or more spans. This lines up with our expectations: whereas an uninformed random sample is very ineffective at surfacing hallucinations, human writers are quite capable of matching our classification; the success of the MAP is contingent on the accuracy of the LVLM-judge.

Inter-annotator agreement. We consider three metrics to assess inter-annotator agreement (IAA).

1. Presence agreement. We measure whether annotators agree on which samples contain hallucinations, regardless of hallucination type and specific span of hallucinated text. We adopt a free-marginal version of Krippendorff's alpha:

$$\alpha_{\text{free}} = 1 - (k/n) \sum_{v,v'}^{V} o_{vv'} \delta(v, v')$$

where n is the total number of paired ratings, k is the number of labels (i.e., 2), δ is the nominal metric, and o is the matrix of observed coincidences. We use the free-marginal formulation because it is less sensitive to differences in label distributions arising from varying rates of hallucinations (Byrt et al., 1993; Randolph, 2005). This property is particularly important as the three sampling strategies are designed to yield different hallucination rates.

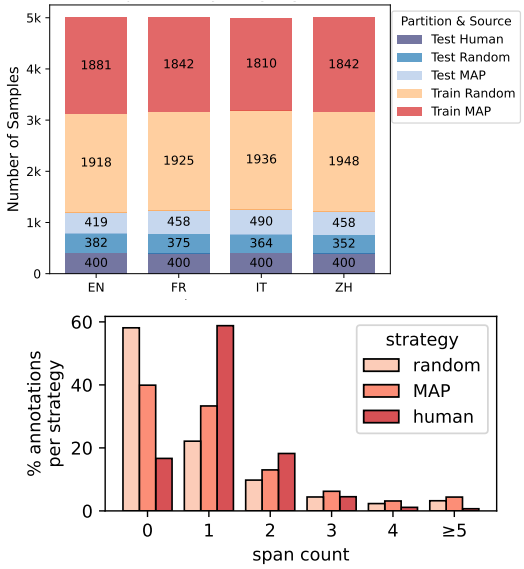


Figure 2: *Top:* Samples per strategy, language, and split. *Bottom:* Proportion of annotations grouped by number of hallucination spans marked in an item.

2. Label agreement. We quantify to what extent annotators agree on the hallucination *type* present in a sample, using a standard (fixed-marginal) Krippendorff's alpha. We contrast agreement across the full dataset (Unconditional) vs. what we observe only on data where all annotators agree on the presence of a hallucination (Given a hallucination).[7]

3. Span agreement. The overlap between spans of text marked as containing hallucinations. We use an Intersection over Union (IoU), as per prior work (Vazquez et al., 2025). We compute the average pairwise IoU across all possible pairs of annotations of the same item. We compute span agreement when factoring in hallucination type in the intersection (Requiring label match) and regardless of assigned label (Ignoring labels).

IAA scores per sampling strategy are reported in Figure 3 (left). Unsurprisingly, controlling for similarity on other annotation levels (presence or absence of hallucination, choice of specific labels) improves agreement rates. More interestingly, are the clear distinctions across strategies: annotators

[7]To handle annotations with more than one span, we ignore all spans beyond the first. See Section C.1 for discussion.

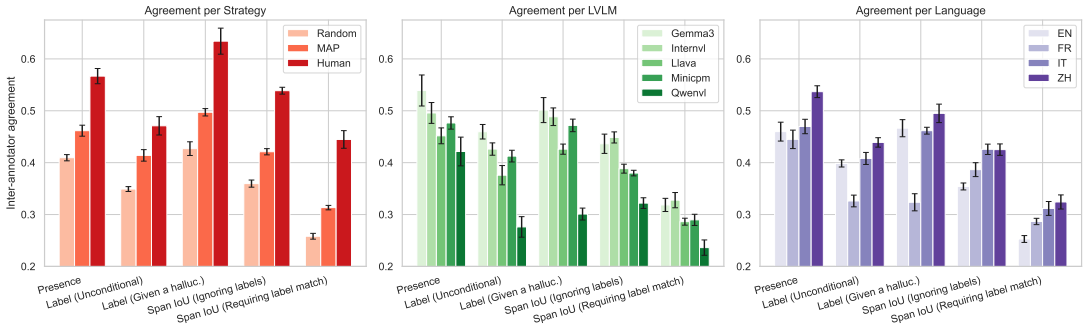


Figure 3: Annotator agreement grouped by sampling process (left), generating LVLM (center), and output language (right). Center and right plots show only LVLM-generated data; left also includes human-generated data. Error bars represent 95% bootstrap confidence, sampling with replacement. See Section C.1 for table versions.

are most consistent on the human-written samples and least consistent on the randomly sampled datapoints, suggesting that sampling strategies very clearly influence the ability of annotators to reach a consensus, and that agreement improves as the selection becomes more controlled.

A key question to assess whether human-written data can serve as a proxy for model-dependent hallucination benchmarks is how IAA varies across generating LVLMs and languages. Such a substitution is only meaningful if the resulting annotations exhibit levels of agreement comparable to the variation naturally induced by changes in model or language. To this end, Figure 3 reports IAA on the random and MAP subsets across LVLM (center) and language (right). None of the generating LVLMs in Figure 3 (center) yields agreement levels higher than those observed for the human-written samples in Figure 3 (left); and variation across generating models is often as large as, or larger than, the variation induced by the sampling strategies themselves. Among the evaluated models, annotators tend to agree most on samples generated by the largest model (Gemma3), whereas QwenVL consistently produces lower agreement scores. Similarly, across the four languages in Figure 3 (right), we observe a similar pattern: agreement varies substantially, with French consistently exhibiting lower IAA and Chinese yielding the highest.

Overall, these results suggest that the increase in agreement observed for human-written samples falls within the range of variation induced by changes in the generating LVLM or the language.

Agreement with expected labels. Do annotators agree with the MAP-judge assigned hallucination labels or the human writers of hallucinations? Figure 4 quantifies the extent to which annotations

3-way classification

Origin	EN	FR	IT	ZH
LVLM-visage	**0.958**	**0.862**	0.859	**0.857**
LVLM-haloq.	0.878	0.847	**0.869**	0.757
human	0.867	0.743	0.738	0.770

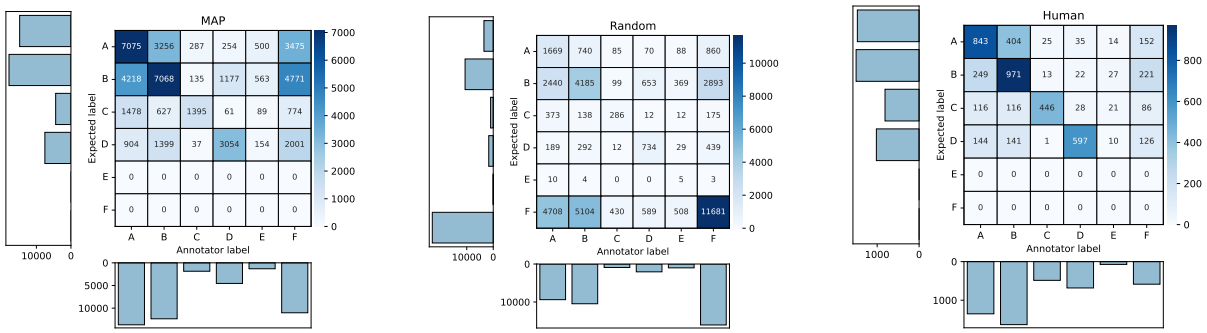
6-way classification

Origin	EN	FR	IT	ZH
gemma3	0.857	0.696	0.595	0.824
internvl	0.429	0.477	0.495	0.474
llava	0.718	0.615	0.654	0.703
minicpm	0.721	0.633	0.655	0.614
qwenvl	**0.877**	**0.832**	0.739	**0.864**
human	0.876	0.748	**0.760**	0.756

Table 3: Per-class F1 scores for 3-way and 6-way classification across languages.

agree with these expected labels. The main departures from agreement along the diagonal are due to two patterns: first, confusion between *invention* and *mischaracterization*, particularly in the MAP and human-written subsets, and secondly, human annotators disagreeing with the labels provided by the MAP and human writers (F columns).[8] We also observe a high false positive rate in terms of binary detection for our LVLM-judge, much higher than what we observe for human-written samples. In contrast, the random data contains a large proportion of correctly identified non-hallucinations; however the MAP judge labels also miss many hallucinations found by annotators (F row). We see that the human-written sampling strategy is more effective at targeting specific hallucination types, since it avoids the high false negative and false positive rates observed when using the MAP strategy with LVLM-judge labels.

[8]Disagreement between writers and annotators is to be expected, given the IAAs reported in Figure 3: what counts as a hallucination is oftentimes subjective (Mickus et al., 2024). This also underscores there are limits in terms of subtlety in judgment we can expect from annotators.



(a) MAP judge labels vs. annotator labels, *MAP* data.

(b) MAP judge labels vs. annotator labels, *random* data.

(c) Writer labels vs. annotator labels, *human* data.

Figure 4: Distribution of expected labels (row) vs. annotator labels (columns) for all available annotations.

Distributional shifts. To assess distributional shift between human-written and LVLM-generated data, we train a logistic regression classifier using character- and word-level n-gram features to predict the origin of each sample in multi-way settings, and report 5-fold cross-validation results in Table 3.

In the **3-way setting** (human-written, LVLM responses on VISaGE and HaloQuest), human-written samples are usually distinguishable from generated data, suggesting a clear distributional shift. However, this separability is comparable to that observed between the two LVLM benchmarks themselves (VISaGE vs. HaloQuest). In the **6-way setting**, where responses are grouped by generating LVLM, human-written data remains identifiable but is not uniquely distinct: Gemma3 and QwenVL are equally or even more easily separable based on textual features alone. Both results suggest that stylistic and lexical variation across LVLMs is already substantial, and that the distributional gap between human-written and LVLM-generated samples is comparable to variation across LVLMs.

5.2 Usefulness as a benchmark

Thus far, we have discussed whether human-written samples have characteristics that would differ significantly from those of LVLM-generated items. We have yet to assess the usefulness of human-written samples when it comes to benchmarking performances on LVLM hallucination detection. We evaluate a modern hallucination detection model, HalluShift++ (Nath et al., 2025), as well as our Gemma3 LLM-judge on our data.

Label assessment with Gemma3 judge. Our first baseline is the Gemma3 LVLM we used earlier as part of the preselection MAP strategy. We now use this model to label items in the random and human-written samples with the six possible labels in Table 1: note that the task here is to label items, not spans. For evaluation, we mark the item as correct if at least one annotator provided the same label as predicted by Gemma3. The results are shown Table 4, broken down by item language and item generation type (model or human), with full results provided in Section C.2. Overall, we see that Gemma3 labels the human-written samples with the highest accuracy, compared to the models; somewhat surprisingly, this holds true even when Gemma3 is the source model. However, human-written samples are not completely outside the range of model samples, especially when looking at F1, precision, or precall rather than accuracy. Note that we *expect* accuracy to shift if detectors are not equally capable on all hallucination types.

		human	gemma3	internvl	llava	minicpm	qwenvl
all	**acc**	0.664	0.543	0.532	0.574	0.485	0.459
	F1	0.499	0.442	0.401	0.455	0.410	0.427
EN	**acc**	0.695	0.522	0.516	0.568	0.449	0.402
	F1	0.511	0.391	0.328	0.516	0.345	0.307
FR	**acc**	0.678	0.539	0.463	0.538	0.481	0.427
	F1	0.509	0.450	0.321	0.449	0.421	0.458
IT	**acc**	0.650	0.539	0.618	0.676	0.519	0.433
	F1	0.482	0.452	0.383	0.516	0.437	0.458
ZH	**acc**	0.635	0.566	0.532	0.508	0.493	0.594
	F1	0.491	0.506	0.505	0.390	0.393	0.425

Table 4: Gemma3 classification: F1 and accuracy.

	gemma3	internvl	llava	minicpm	qwenvl
human	0.796	0.770	0.916	0.867	0.708
gemma3		0.717	0.771	0.700	0.769
internvl			0.658	0.800	0.660
llava				0.831	0.683
minicpm					0.766

Table 5: Comparison of performance rankings for human- and machine written data (Pearson's r, $n = 40$).

This illustrates that the error profile is roughly conserved on the human-written data, but it still leaves open one major point: whether performance rankings from different judges would be accurately captured by our human-written data. To that end, we select ten LVLMs to serve as judges (Gemma 3 4b, 12b and 27b, internVL 4b, 8b and 38b, QwenVL 2b, 4b, 8b and 32b) and compare their performances on human-written vs machine-written data. For each judge, we compute F1 scores per language (i.e., generated by a specific LVLM or by human writers), for a total of 40 setups per data source. We then measure whether performances are correlated across sources, using a Pearson correlation coefficient. This yields the scores displayed in Table 5: crucially, we find that human-written data yields the highest correlations. In other words, human-written samples are as adequate as samples generated by any arbitrarily chosen LVLM.

Labeled-span assessment with HalluShift++. Next, we consider a baseline inspired by HalluShift++, which detects hallucinations from internal LVLM signals, including representation shifts, attention patterns, and confidence-related features. We adapt this idea to our span-level localization setting by running a teacher-forced forward pass over each image–question–answer triple and extracting token-level features from an evaluator model. We extract these features from Llava-NeXT, yielding both a model-dependent self-probing setup as well as an external evaluation setup, depending on which LVLM generated the samples in our data. The features are used to train a token-level classifier in a multiclass setting where tokens are assigned to one of the labels A–F (Table 1).[9]

As Table 6 shows, usually, we obtain low F1 scores, driven primarily by a low precision, highlighting the challenging nature of our collected data. Importantly, performances on the human-written data (bottom row of Table 6a) are comparable to what we observe from specific generating LVLMs.

One of the reasons for the modest performances we observe in Table 6 is the high prevalence of non-hallucinated tokens. This prompts us to recalibrate model probabilities for non-hallucination labels by computing the optimal probability threshold in a one-versus-all setting, comparing F to all other labels. We calibrate on data derived from all possible origins (human data or any of the 5 LVLMs) and record the delta in F1 in Table 7. This usually leads to improvements, with one exception: calibration on Llava systematically degrades performances. As this LVLM is also the backbone for the detection pipeline, this appears to illustrate the biases that model-dependent benchmarks can introduce.

[9]More details and experiments, including with EUQ (Huang et al., 2026), are provided in Sections C.3 and C.4.

Strategy	acc	prec	rec	F1
random	0.440	0.226	0.506	0.204
MAP	0.403	0.240	0.517	0.224
human	0.332	0.240	0.443	0.203

(a) HalluShift++ classification metrics across strategies.

	random	MAP	all
gemma3	0.157	0.227	0.189
internvl	0.163	0.203	0.186
llava	0.186	0.224	0.212
minicpm	0.182	0.205	0.195
qwenvl	0.225	0.209	0.221

(b) HalluShift++ F1 per generating LVLM.

Table 6: HalluShift++ performance.

		calibrate on					
		human	**gemma3**	**internvl**	**llava**	**minicpm**	**qwenvl**
test on	**human**	0.064	0.056	0.053	−0.012	0.050	0.070
	gemma3	0.070	0.069	0.065	−0.009	0.062	0.077
	internvl	0.043	0.040	0.039	−0.008	0.038	0.044
	llava	0.046	0.047	0.047	−0.007	0.047	0.041
	minicpm	0.048	0.047	0.046	−0.007	0.045	0.047
	qwenvl	0.062	0.061	0.060	−0.008	0.059	0.063

Table 7: F1 delta after recalibration (all data).

6 Discussion

The core question we address in this work is whether human-written samples can be used to evaluate hallucination detection models. To this end, we have constructed a large dataset of vision-language hallucination samples using three sampling strategies: random sampling, model-assisted pre-selection, and human-written items. Our analyses provide evidence that human-written samples are a viable basis for evaluating hallucination detection. First, they allow us to more carefully control the actual contents of our dataset (Figure 4). Second, annotators reach a higher degree of consensus on these items (Figure 3). Third, human-written samples exhibit a distinct but comparable distributional profile relative to LVLM responses (Table 3), and the performance of hallucination detectors on human-written samples is broadly consistent with their performance on LVLM-generated data, while avoiding pitfalls that model-dependent benchmarks suffer from (Table 7). It is worth stressing that the

second of these points is not trivial: prior work has frequently highlighted that annotators express genuine disagreement as to what counts as a hallucination (Mickus et al., 2024) and where hallucination spans begin and end (Vazquez et al., 2025; Schmidtova et al., 2026). While in general, some of this discrepancy can be imputed to gaps in expertise (Calò et al., 2026), our experimental setup allows us to cleanly isolate the effects of data sampling since the same cohort of annotators and the same inputs were used throughout our dataset.

Our work is not without its caveats. First, we observe that the data we collect is remarkably challenging, with some models performing at or close to random on specific subsets. Second, we cover a small number of hallucination detection models; establishing whether performance on human-written samples reliably predicts performance on true hallucinations requires broader empirical validation. Results in Table 5 are encouraging, but stronger guarantees will require concerted research efforts. A shared task is currently underway to further validate this point.[10] A similar comment holds for the representativeness of data derived from different LVLMs, as we do not systematically quantify variation across all possible generators. At the same time, human-written samples can always serve as a sanity check for hallucination detection systems: because they are constructed to contain the target phenomena, a large performance gap between human-written and LVLM-generated data would likely indicate reliance on spurious signals. Our approach alleviates some of the issues that plague modern NLP evaluation, yet it also frames hallucination detection as a black-box problem. Methods relying on internal model signals will require adaptation, as discussed in Section C.3 for Hallushift++. This also reflects a broader trend toward proprietary black-box models, where evaluation must proceed without access to internal states.

It is worth acknowledging that there can be concerns as to the fitness of human-written data for evaluation purposes, as humans could introduce their own biases.[11] Previous work has found genuine mismatches between human and system-based assessments, especially on disagreement modeling (e.g., Pavlick and Kwiatkowski, 2019; Mickus et al., 2025). However, such mismatches appear to exhibit a profile distinct from that observed in the present work: the gaps described in these studies appear to emerge between human and machine assessments, whereas what we target (and document in Table 7) is bias specific to one particular model. More broadly, prior work has documented socionormative and cognitive biases of annotators (Mieleszczenko-Kowszewicz et al., 2023; Gautam and Srinath, 2024; Blodgett et al., 2020) and how they can impact performances (Geva et al., 2019). The biases observed in NLP systems tend to be exacerbated versions of human socionormative biases (Sheng et al., 2021; Chen et al., 2024a; Almatarneh et al., 2025), i.e., prior work suggests that human annotators present *milder* cases of socionormative biases. If socionormative biases can serve as a spurious, exploitable feature that can compromise evaluation, then we should therefore expect human data to be preferable to data generated by NLP systems. As far as SHEEP is concerned, our dataset construction sidesteps some of the issues pointed out in prior work. In particular, we keep track of annotators, hence future studies can follow Geva et al. and Gautam and Srinath's recommendations.

7 Conclusion

We present SHEEP, a dataset of 20,000 items across four languages (Chinese, English, French, Italian) for hallucination detection across 5 LVLMs. We assess three strategies for obtaining samples: randomly selecting outputs of LVLMs, using an LVLM-as-an-annotator pre-selection, and writing samples manually. The dataset is richly annotated with 5 possible hallucination labels at the span level, with each item assessed independently by three annotators. The performance of modern hallucination detectors underscores the challenging nature of SHEEP. We find that human-written samples exhibit several beneficial characteristics: they yield items that are more consensual among our annotators while maintaining a distributional profile similar to that observed in LVLM-generated samples. We argue that this strategy avoids some pitfalls common in hallucination evaluation, including a tendency toward rapid obsolescence as models quickly cease to represent the state of the field.

[10] See helsinki-nlp.github.io/shroom/2026/

[11] Bias is a valence-heavy and ambiguous word that can refer to a wide variety of phenomena — ranging from judgments made on socionormative grounds, to distributional and structural differences between data sources (in this case, human-written vs machine-written). As far as the latter goes, we argue that *some* biases can be good: e.g., it is genuinely beneficial for us to produce more instances of OCR and miscounting, as long as we do not introduce trivial distributional patterns.

Limitations

We frame this study as a first outlook into whether human-written data could replace LVLM-generated data for hallucination detection benchmarking. As such, there are clear limitations due to the exploratory nature of this project: The tasks and setups considered here are not an exhaustive or representative subset of all applications and contexts where LVLM hallucination detection is relevant; we focus primarily on high-resource languages, although prior literature shows that hallucinations do not impact all languages equally (Vazquez et al., 2025; Gamba et al., 2026; Datta et al., 2026).

An important consideration is that we propose a static dataset, hence it is in principle possible for leakage to occur. As of writing, our testset labels are kept private and accessible through a dedicated scoring interface at `https://shroom.pythonanywhere.com/` so as to mitigate this risk.

Ethical Considerations

Data collection risks. The present work describes a data collection process, and therefore several ethical considerations apply. The data was deemed unlikely to present any potential harm to annotators, as inputs strictly target common objects with uncommon attributes (VISaGE, Frank and Allaway, 2025) or commonplace questions and images mismatched with one another (HaloQuest, Wang et al., 2024b). The dataset contains no offensive or harmful topics, nor any personally identifying information. Annotators' usernames on the annotation platform are anonymized in the final distributed dataset.

Potential risks of the research agenda. This work targets hallucination evaluation in LVLM technology. We believe that fostering a research ecosystem that actively combats the spread of misinformation and technologies that can produce misinformation at scale is of great importance. While part of this ecosystem has to include evaluation benchmarks such as the one we propose, we recognize that by design our dataset contains information that is false, incoherent, or misleading, and strictly emphasize that the intended use of this artifact is to promote research into hallucination detection and mitigation.

Acknowledgments

The construction of this dataset was made possible thanks to a grant from the Finnish Society of Sciences and Letters. This work has also received funding from the Digital Europe Programme under grant agreement No 101195233 (OpenEuroLLM). The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the EU.

References

Sattam Almatarneh, Ahmed BaniMustafa, Ghassan Samara, Raed Alazaidah, and Qasem Obeidat. 2025. Bias and fairness in NLP: Addressing social and cultural biases. In *Advances in Computational Intelligence: 18th International Work-Conference on Artificial Neural Networks, IWANN 2025, A Coruña, Spain, June 16–18, 2025, Proceedings, Part II*, page 609–619, Berlin, Heidelberg. Springer-Verlag.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

T Byrt, J Bishop, and J B Carlin. 1993. Bias, prevalence and kappa. *J Clin Epidemiol*, 46(5):423–429.

Eduardo Calò, Saad Mahamood, Albert Gatt, and Kees Van Deemter. 2026. A logic-based approach to hallucinations in data-to-text NLG: Experiments with human and LLM annotators. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 28–62, San Diego, California, United States. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Debtanu Datta, Mohan Kishore Chilukuri, Yash Kumar, Saptarshi Ghosh, and Muhammad Bilal Zafar. 2026. Do LLM hallucination detectors suffer from low-resource effect? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2959–2985, Rabat, Morocco. Association for Computational Linguistics.

Arthur P Dempster. 1967. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54(3-4):515–528.

Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. 2024. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 10707–10715, New York, NY, USA. Association for Computing Machinery.

Stella Frank and Emily Allaway. 2025. VISaGE: Understanding visual generics and exceptions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32549–32558, Suzhou, China. Association for Computational Linguistics.

Federica Gamba, Aman Sinha, Timothee Mickus, Raul Vazquez, Patanjali Bhamidipati, Claudio Savelli, Ahana Chattopadhyay, Laura A. Zanella, Yash Kankanampati, Binesh Arakkal Remesh, Aryan Ashok Chandramania, Rohit Agarwal, Chuyuan Li, Ioana Buhnila, and Radhika Mamidi. 2026. Confabulations from acl publications (cap): A dataset for scientific hallucination detection. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2512–2524, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in NLP. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 82–88, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Tao Huang, Rui Wang, Xiaofei Liu, Yi Qin, Li Duan, and Liping Jing. 2026. Detecting misbehaviors of large vision-language models by evidential uncertainty quantification. *arXiv preprint arXiv:2602.05535*.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 525–534, New York, NY, USA. Association for Computing Machinery.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun

Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2025. Datasets for large language models: a comprehensive survey. *Artif. Intell. Rev.*, 58(12).

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 37–58, Bangkok, Thailand. Association for Computational Linguistics.

Timothee Mickus, Aman Sinha, and Raúl Vázquez. 2025. Your model is overconfident, and other lies we tell ourselves. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5401–5417, Vienna, Austria. Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanisław Woźniak, Ewa Dzięcioł, Przemyslaw Kazienko, and Jan Kocoń. 2023. Capturing human perspectives in NLP: Questionnaires, annotations, and biases. In *Proceedings of the Second Workshop on Perspectivist Approaches to NLP (NLPerspectives)*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Sujoy Nath, Arkaprabha Basu, Sharanya Dasgupta, and Swagatam Das. 2025. Hallushift++: Bridging language and vision through internal representation shifts for hierarchical hallucinations in mllms. *arXiv preprint arXiv:2512.07687*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. HalLoc: Token-level localization of hallucinations for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29893–29903.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. 2024. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357, Mexico City, Mexico. Association for Computational Linguistics.

Nhi Pham and Michael Schott. 2024. H-pope: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. *Preprint*, arXiv:2411.04077.

Qwen Team. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qwen Team. 2025b. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k[free]): An alternative to fleiss' fixed-marginal multirater kappa. Paper presented at the Joensuu Learning and Instruction Symposium.

Anku Rani, Vipula Rawte, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. 2024. Visual hallucination: Definition, quantification, and prescriptive remediations. *Preprint*, arXiv:2403.17306.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Kuniaki Saito, Risa Shinoda, Shohei Tanaka, Tosho Hirasawa, Fumio Okura, and Yoshitaka Ushiku. 2026. Haldec-bench: Benchmarking hallucination detector in image captioning. *Preprint*, arXiv:2511.20515.

Patricia Schmidtova, Ondrej Dusek, and Saad Mahamood. 2026. Hotelcheckspan: A benchmark dataset for llm faithfulness. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9973–9987, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024b. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Computer Vision – ECCV 2024*, pages 288–304, Cham. Springer Nature Switzerland.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, and 1 others. 2024. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8395–8419.

Bei Yan, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. 2026. Measuring the measurers: Quality evaluation of hallucination benchmarks for large vision-language models. *Preprint*, arXiv:2406.17115.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. 2024. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, and 15 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *Preprint*, arXiv:2509.18154.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Contributions of individual authors

Timothee Mickus: overall organization, experimental designs, annotator recruitment, inter-annotator agreement (free-marginal for presence agreement), span annotation platform development, guideline development, data selection, data creation (English).
Claudio Savelli: HalluShift++ implementation, translation pipeline development, code development, guideline development, data creation (Italian).
Eduardo Calò: Theoretical framework, state of the art surveying, guideline development, data creation (Italian).
Emilio Raimond: LLM-as-a-judge implementation, code development, state of the art surveying, data creation (French).
Stella Frank: Guideline development, data selection, data creation (English), experimental design.
Hengyu Luo: Data creation (Chinese), translation pipeline development, code development,
Flavio Giobergia: Inter-annotator agreement, data quality checking.
Vincent Segonne: Data creation (French), translation pipeline development, sample-writing platform development.
Chuyuan Li: Distributional shift analysis.
Aman Sinha: EUQ implementation.
Lorenzo Vaiani: Code development.
Jörg Tiedemann: Organizational support.
Raúl Vázquez: overall organization, experimental designs, annotator recruitment, annotator briefing, span annotation platform development, guideline development, data selection.

B Reproducibility details

B.1 Translation pipeline

Model selection. We evaluated candidate translation models by manually assessing 100 randomly sampled instances per target language (Italian, French, and Chinese) to measure adequacy and fluency. Based on this evaluation, we adopt a hybrid strategy. **NLLB-200 (3.3B)** (NLLB Team et al., 2022) is used for translations between English and Romance languages (Italian and French), where it performs reliably. For Chinese, we instead use **Qwen3 8B** (Qwen Team, 2025a), which we found produces more fluent, semantically accurate outputs, particularly for short prompts and question-like inputs.

Translation settings. We apply the translation pipeline in two scenarios: (i) **English prompt translation**, where all prompts and references are translated from English into the other three languages, as both datasets are originally in English solely. (ii) **multilingual augmentation**, used for human-written samples, where inputs may originate in any of the four languages and are translated into the remaining three. All these samples are manually checked and post-edited when necessary, ensuring high-quality multilingual consistency.

In the second setting, Qwen3 is used for all translation directions involving Chinese ($\leftrightarrow$), while NLLB is used for directions that do not involve Chinese (e.g., Italian $\leftrightarrow$ French or English).

NLLB pipeline. For Italian and French (and more generally for translations not involving Chinese), we use a sequence-to-sequence pipeline based on NLLB-200. We explicitly specify source and target language codes (e.g., `eng_Latn` $\rightarrow$ `ita_Latn`) and enforce the target language via a forced beginning-of-sequence token. Decoding is performed using beam search with 4 beams and no sampling.

Qwen pipeline. For Chinese and all translation directions involving Chinese, we use Qwen3 8B in an instruction-following setup. We design a structured prompt that enforces strict translation behavior (i.e., no explanations, no additional text, and no question answering). The full prompt template is reported in Figure 5. We use the tokenizer chat template and perform decoding with beam search (4 beams) and no sampling to ensure deterministic outputs.

B.2 Gemma3 model-assisted pre-selection prompt

We provide the exact prompts used in our experiments in Figure 6. Because the purpose of the judge is to support large-scale pre-selection rather than to provide final annotations, we ask it to output only the predicted label, without a reasoning trace.

B.3 Guidelines & Annotation interfaces

Guidelines for annotators and writers. In this appendix, we include a copy of the guidelines provided to human writers (Figure 7) and annotators (Figures 8 and 9). The annotators had access to their guideline as a single document (split across

```
You are a translation engine. Translate only. Never explain, never reason, never add notes. Output only the final
translation text.
```

(a) System prompt.

```
You are a translation engine. Translate SOURCE_TEXT into TARGET_LANGUAGE.
Rules:
1) Translate exactly what is written.
2) If SOURCE_TEXT is a question, translate the question; do NOT answer it.
3) Keep named entities, numbers, URLs, and facts unchanged unless translation requires adaptation.
4) Output only the translated text.
5) Do not add explanations, notes, or extra lines.
SOURCE_TEXT: <input text>
```

(b) User prompt template.

Figure 5: Prompt template used for Qwen3 translation. `TARGET_LANGUAGE` is replaced with the desired language (e.g., Italian, French, Chinese), and `<input text>` with the source string.

two pages in the present reproductions for pagination purposes); minor details such as a table of contents have been omitted. Annotators also had access to a user guide for the annotation platform they were required to use, which is not reproduced here.

Annotation interface We developed a bespoke annotation interface using a stack composed of Django 5, jQuery and Bootstrap, deployed with NGINX and Gunicorn. The platform was hosted on a server and used by annotators both for the training phase and for the actual annotation experiment. Figure 10 provides a screenshot of the annotation interface for an experimental item, including the image, question, and answer. Annotators could highlight spans directly in the answer text. Upon highlighting a span, a box appeared asking them to assign one label to the selected span (Figure 11). Annotated spans were color-coded according to their assigned labels (Figure 12). Annotators were also free to leave additional remarks in a dedicated comment box. A similar platform with minor modifications was used to collect human-written samples, as shown in Figure 13.

C Supplementary results

C.1 Detailed results for IAA assessments

In Table 9, we present IAA measurements per strategy; in Table 10, per language and model. These correspond to Figure 3 (left) and Figure 3 (center, right) respectively.

Effect of label aggregation process on IAA. In Section 5.1, we mention that, when computing label agreement, we aggregate observations from annotators by simply selecting the label of the first span provided by an annotator. An alternative strategy consists of selecting the label of the longest span instead, in terms of character coverage. A comparison of these two strategies is provided in Table 8: as is apparent, the first-span approach yields higher IAA scores than the longest-span strategy; differences are statistically significant under our bootstrapping. It is not entirely clear to us why this is the case, we conjecture that it has to do with hallucination 'snowballing' (Zhang et al., 2025), which annotators might react to in different ways.

C.2 Supplementary results on Gemma3-assisted labeling

Full classification results. In Table 11, we provide a complete overview of classification metrics for the label assessment using the Gemma3 LLM-judge. Similar observations as mentioned in the main text hold.

Gemma3 assessment across languages and generating LVLM. For LLM-based external labels, we additionally study the distribution of labels assigned by the Gemma3 judge across models and languages, as shown in Figure 14. While the distribution of LLM-based assessment across languages is stable, noteworthy patterns emerge when comparing models: Gemma3 rates its output as less likely to contain hallucinations, whereas Internvl and Llava are rated as containing more mischaracterizations.

```
Your role is to assign a label to the provided text and image based on whether or not the text contains a
*hallucination*.
Hallucinations correspond to cases where the provided text states information that contradicts the image.

LABELING INSTRUCTIONS:
You must assign one of the following labels (A, B, C, D, E or F):

A. **Invention.** The text refers to an entity, object, property, or event that is not present in the image at
all. This includes but is not limited to: fabricated objects, invented people, imagined actions unsupported by the
image
Examples:
- "red bus" (no bus visible in the image)
- "a dog nearby" (no dog visible in the image)
- "an umbrella" (not present in the image)

B. **Mischaracterization.** The text refers to something that is present in the image but describes it incorrectly.
This includes but is not limited to: wrong object type (e.g., a taxi described as a bus), wrong colors (e.g., a
blue shirt described as green), an incorrect identity (e.g., a person standing described as sitting).
Examples:
- Calling a van a "bus"
- Describing a blue shirt as "green"
- Describing a dog running when the animal is standing or laying down.

C. **Wrong Reading.** The hallucination arises from misreading text that is visible in the image. This applies
whenever there is something readable in the image itself, but the provided text misquotes it, alters it, or
misunderstands it.
Examples:
- the image includes a "STOP" sign but the provided text mentions a "SHOP"
- the image includes a "50% OFF" sign but the provided text mentions "30% OFF"

D. **Miscounting.** The text expresses an incorrect quantity of visible items. This includes but is not limited
to: wrong counts of objects shown in the image (e.g., people, cars, animals...); incorrect quantities; explicit
numeric misstatements....
Examples:
- the text refers to "three elephants" when only two elephants are visible in the image.
- the text refers to "many people" when only two persons are visible in the image

E. **Other.** A hallucination that does not fit categories A, B, C or D.

F. **No hallucination.** The provided text does not contain any hallucination: the text does not contradict the
image and is perfectly coherent with the information shown in the image.

Do not explain your reasoning. Only reply with the letter corresponding to the label you assign, i.e., your answer
must contain only the letter A, B, C, D, E or F.

PROVIDED TEXT:
```

Figure 6: Prompt used for the LLM-judge (Gemma3 27B IT).

This document outlines guidelines for creating samples for the human-written section of DATASET.

Motivation
One of the things we want to verify with DATASET is whether we can evaluate the quality of hallucination detectors independently from any specific LLM output. To that end we intend to include a human-written subset of outputs in the test set, so as to be able to compare performances on LLM-generated and human-written hallucinations.

Overview of the task
You will be writing answers to a series of provided questions grounded on images. You will need to imitate the style of LLMs and make sure that the answer you provide contains a hallucination. Read these guidelines carefully and in their entirety before you start

You are not allowed to use an LLM to generate your answers. Answers should be written from scratch.

All other rules are provided as guidance: use your best judgment, and leave comments on this document whenever relevant.

Detailed instructions

- Before you start, read the guidelines for annotators to keep in mind how the data will be presented to annotators and what task they will need to perform:
  <URL>
  this document also contains the following items, which you will need to be familiar with:
  - a definition of what hallucinations are
  - a taxonomy of possible types of hallucinations (labels A, B, C, D or E)
- Written samples are to be collected through the dedicated annotation platform, available from our private github repo: <URL>
- Your written samples should be plausible answers to the questions as provided. Do not modify the questions.
- Each answer you write should contain a mistake corresponding to one of the four specific labels A, B, C or D (but not E). For instance, if your answer contains a mistake that pertains to an OCR problem (label C), it should not also include a made up entity (label A)
- Your answer must otherwise be coherent, grammatical, and self-consistent.
- Across all answers that you produce, you should target a balanced proportion of all four labels (25% of A, 25% of B, 25% of C, and 25% of D) Your mistakes must be discoverable given the image. The annotator should not have to perform an external google search to spot the mistake you included.

In terms of style:

- Before you start writing samples, familiarize yourself with LLM style and the type of hallucinations they make. You can use the self-hosted pilot interface available at: <URL>
- Do not look at samples written by other organizers, as much as possible.
- Produce diverse samples.
  - The mistakes you include do not have to target the most salient part of the question or image. For instance, consider including mistakes that target
    - objects in the background,
    - spatial relationships between objects
    - superfluous claims not directly/immediately relevant to the question prompt
  - Hallucinations can often be triggered by surprising images that depict non-standard or otherwise exceptional situations (such as three-legged cats, red bananas, cows wearing Christmas trees on their heads. . .), in which case LLMs can either fabricate completely ungrounded information or generate text corresponding to the most probable situation (e.g., “this cat has four legs”). Consider whether the image provided could confuse an LLM.
  - Consider imitating different styles of LLMs, such as
    - samples with or without markdown
    - samples with or without a sycophantic opening line (“excellent question!”)
    - samples with or without explicit markers of reasoning (“To answer this question, I will carefully analyze all components of the image”).
  - Your answers can be of varying length, from a single sentence to a few paragraphs. They should not be overly short, since this would indicate that there is not much in the answer aside from the hallucination.

Figure 7: Guidelines for human-written samples production. Actual URLs are replaced with the placeholder <URL>.

Introduction

In this annotation project you will be shown a series of question-answer pairs, plus a relevant image. The answer will be a passage of text produced by a Large Language Model (LLM) in response to the question. You will be asked to identify, with respect to the image, which spans of text in the answer constitute a "hallucination", i.e., which parts of the answer are not supported by the image, given the question.

Each datapoint consists of:

- An image
- A question related to the image (i.e., a prompt)
- A response produced by a model

General Instructions

1. Carefully examine the image, and read the prompt, and model response.
2. Highlight each span of text in the model response that is not supported by the information in the image (i.e., contains a hallucination or overgeneration). Base this exclusively on the definition of hallucination we are using (see below).
   Important:
   - Annotate conservatively: annotations must include only the minimum number of characters that would need to be edited or deleted in order to make the answer correct.
   - Prefer highlighting content words over function words.
   - Avoid selecting full sentences unless absolutely necessary.

   These are not rigid rules, i.e., use your best judgment when needed.
3. Classify each marked span using the proposed taxonomy (see below). Each span must be marked with only one class label (A, B, C, D, or E).
4. If a response contains multiple hallucinations:
   - Highlight each span separately.
   - Assign one label (A–E) per span. Do not merge unrelated errors into a single span.
   - Do not overlap or embed spans.
5. If a response contains no hallucination:
   - write "No hallucination" (in English) in the comment box.

You can also use the comment box to mark anything else that you want to raise to our attention, such as:

- model answers that are ungrammatical,
- outputs that are generated in the wrong language,
- cases where you are especially uncertain,
- anything else you deem relevant.

Please write comments in English, so that the whole team can read them.

What counts as a hallucination?

Definition. Content that contains or describes facts that are not supported by the image. In other words: hallucinations are cases where the answer text is inaccurate, or more specific than it should be, given the provided image.

A hallucination is any content that:

- Refers to something not present in the image
- Misdescribes something visible in the image
- Incorrectly counts visible entities
- Misreads visible text
- Or otherwise goes beyond what can be visually grounded

What to do with unverifiable statements & expressions of uncertainty in model outputs?

Note that some model responses introduce information that cannot be verified from the image alone. These are statements that add external facts, background explanations, or narrative details that cannot be inferred directly from the image.

If the statement introduces new factual information that cannot be verified from the image, it should generally be treated as a hallucination. For example:

- Hallucination to annotate: statements of factual information not visible in the image ("This is a beach in Thailand" for a generic-looking tropical beach).
- Plausible visual inference: do not annotate: "This is a street in Paris" when the Eiffel Tower is in the background.

Model responses may also include uncertainty-marking language (e.g., "it looks like", "it seems", "possibly", "it might be"). In such cases, do not annotate the uncertainty phrase itself. Instead, evaluate whether the underlying claim is visually supported.

Examples:

- "The man is a teacher." -> highlight "teacher" (unless the image very clearly indicates this is the case, for example if he is standing in front of a classroom of students.)
- "This restaurant serves Mexican food." -> highlight "Mexican" (when the restaurant in the image could serve other types of food, there is no recognizable food in sight nor any other sign of the restaurant being Mexican)
- "He looks sad because he lost his job" in an image with clear visual cues connecting sadness to job loss. -> DO NOT HIGHLIGHT (Not a hallucination: need to consider "sadness" being supported by facial expression and posture, Job loss is visually grounded by: a termination notice, packing office belongings into a box, etc.)

Figure 8: Annotator instructions, part 1 of 2: overview and general definitions.

Proposed taxonomy of hallucinations
This classification consists of 5 classes of hallucinations.

A. Invention
The highlighted span refers to an entity, object, property, or event that is not present in the image at all.
Includes: fabricated objects, invented people, imagined actions unsupported by the image
Examples:
- "red bus" (no bus visible)
- "a dog nearby" (no dog visible)
- "an umbrella" (not present)

Decision rule: If the entity does not exist in the image choose A

B. Identity incongruity (Mischaracterization)
The span refers to something that is present in the image but describes it incorrectly.
Includes:
- Wrong object type (taxi -> bus)
- Wrong color (blue shirt -> green)
- Wrong identity (man -> woman)
- Wrong activity (standing -> sitting)
- Incorrect spatial relations between objects or entities (below -> above, right -> left)

Examples:
- Calling a van a "bus"
- Describing a blue shirt as "green"
- Describing a dog running when the animal is standing or laying down
- Saying a cup is on the table when it is actually in someone's hand
- Saying a person is behind the car when they are standing next to it

Decision rule: If the object exists but is misidentified or misdescribed choose B

C. Wrong Reading / OCR Problem
The hallucination arises from misreading text that is actually visible in the image.
Applies only if there is readable text in the image and the model misreads or alters it.
Examples:
- Image: "STOP" but response has to do with "SHOP"
- Image: "50% OFF" but response is about "30% OFF"

Do NOT use C in case:
- The text does not exist in the image. Use A in this case
- The error is numerical counting (not text reading). Use D in this case

D. Numeric Discrepancy / Miscounting
The span expresses an incorrect quantity of visible items.
Includes:
- Wrong count of people, cars, animals
- Incorrect quantities
- Explicit numeric misstatements

Examples:
- "three people" (only two visible)
- "two dogs" (only one visible)
- "N something" (the thing is uncountable)

Do NOT use D in case: If a number is misread from visible written text. Use C instead.

E. Other (Use Sparingly)
A hallucination that does not fit categories A to D. Try to use as minimally as possible.
When using this tag, please leave a comment indicating why you have used it.

Figure 9: Annotator instructions, part 2 of 2: classification of hallucinations.

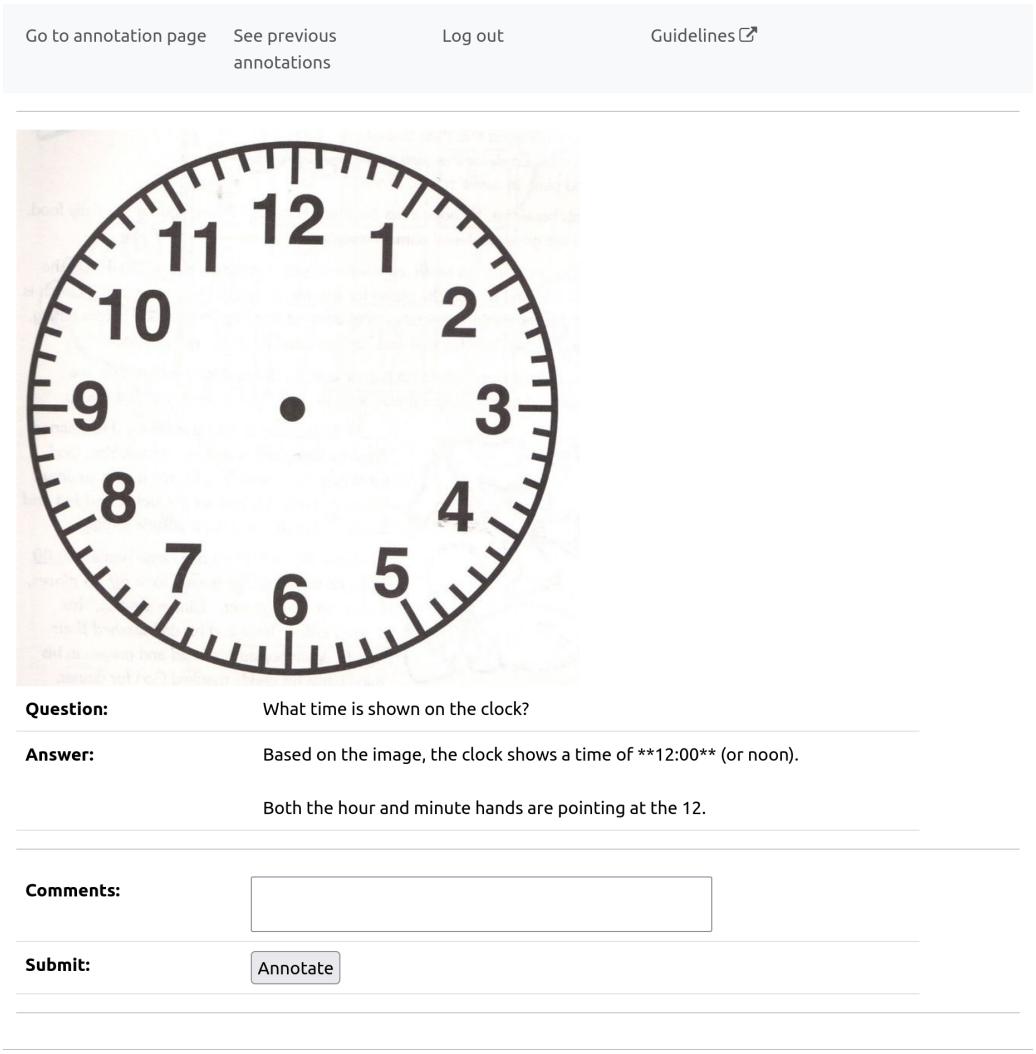


Figure 10: Screenshot of the annotation interface used (experimental item).

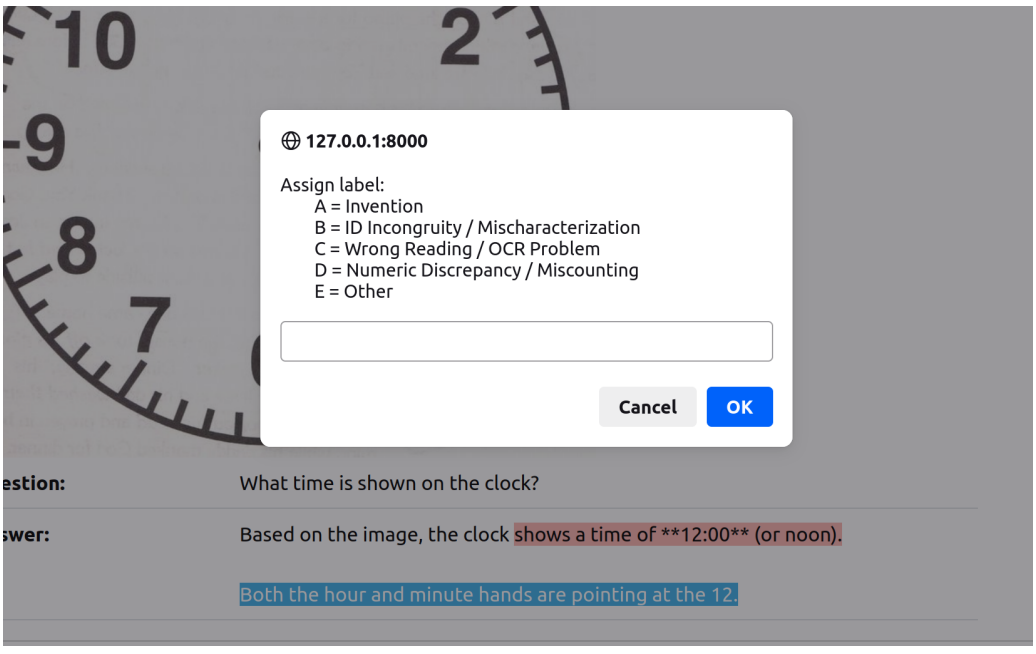


Figure 11: Screenshot of the annotation interface used (hallucination category selection box).

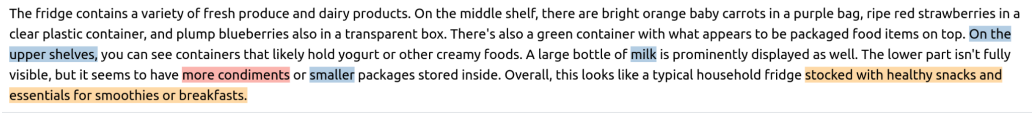


Figure 12: Screenshot of the annotation interface used (hallucination span color-coding).

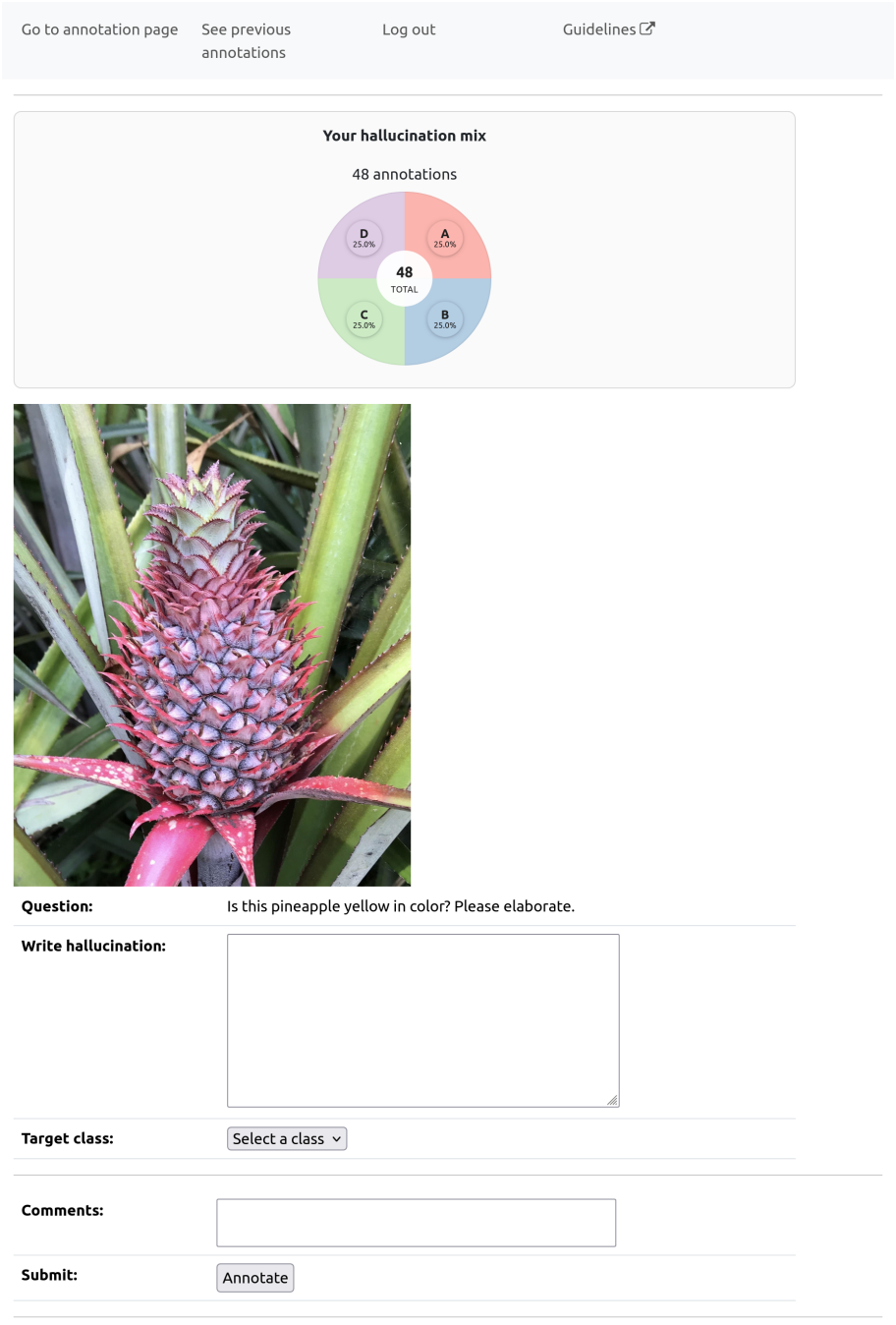


Figure 13: Screenshot of the collection interface used for human-written samples. The pie chart on top of the page provides the distribution of intended labels.

	Label agreement (longest)		Label agreement (first)	
	All labels	**Hallu. only**	**All labels**	**Hallu. only**
Random	0.3346 ± 0.0067	0.3453 ± 0.0088	0.3716 ± 0.0150	0.4328 ± 0.0148
MAP	0.3925 ± 0.0102	0.4126 ± 0.0089	0.4475 ± 0.0141	0.5020 ± 0.0113
Human	0.4395 ± 0.0149	0.4713 ± 0.0163	0.5881 ± 0.0145	0.6318 ± 0.0189
model=gemma3	0.4427 ± 0.0142	0.4611 ± 0.0134	0.5089 ± 0.0198	0.5385 ± 0.0164
model=internvl	0.4114 ± 0.0143	0.4282 ± 0.0142	0.4595 ± 0.0226	0.4866 ± 0.0222
model=llava	0.3419 ± 0.0122	0.3730 ± 0.0127	0.3810 ± 0.0149	0.4463 ± 0.0188
model=minicpm	0.3914 ± 0.0113	0.4049 ± 0.0116	0.4036 ± 0.0201	0.4746 ± 0.0253
model=qwenvl	0.2626 ± 0.0109	0.2786 ± 0.0086	0.3628 ± 0.0205	0.4289 ± 0.0183
language=EN	0.3698 ± 0.0103	0.4033 ± 0.0068	0.3871 ± 0.0111	0.4501 ± 0.0143
language=FR	0.3104 ± 0.0091	0.3204 ± 0.0063	0.4119 ± 0.0252	0.4592 ± 0.0273
language=IT	0.3963 ± 0.0126	0.4090 ± 0.0146	0.4123 ± 0.0205	0.4599 ± 0.0194
language=ZH	0.4221 ± 0.0092	0.4380 ± 0.0116	0.4785 ± 0.0240	0.5360 ± 0.0170

Table 8: Effects of agglomeration policy for multi-span annotations on label annotator agreement for *Model-assisted pre-selected (MAP)*, *random* and *human* data, as well as separately for generating model, and output language (on MAP + random data).

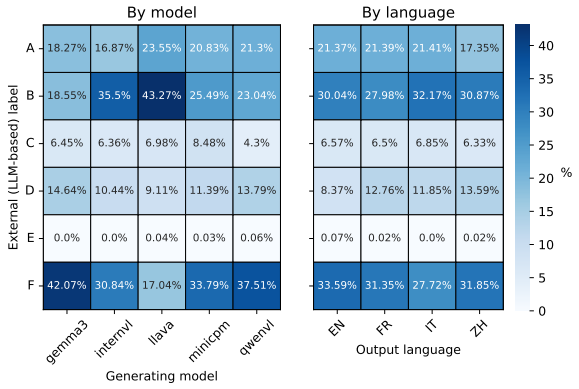


Figure 14: Distribution of labels assigned by an LLM judge (Gemma3).

Agreement Type	Metric	Random	MAP	Human
Pres.		0.41 ± 0.01	0.46 ± 0.01	0.57 ± 0.01
Label	**Unconditional**	0.35 ± 0.00	0.41 ± 0.01	0.47 ± 0.02
	Given a hallu.	0.43 ± 0.01	0.50 ± 0.01	0.63 ± 0.03
Span	**Requiring label match**	0.26 ± 0.01	0.31 ± 0.00	0.44 ± 0.02
	Ignoring labels	0.36 ± 0.01	0.42 ± 0.01	0.54 ± 0.01

Table 9: Annotator agreement for *Random*, *MAP*, and *Human* data. Values reported as mean ± 95% bootstrap confidence interval, sampling items with replacement.

C.3 Supplementary results on HalluShift++

HalluShift++ implementation details. Our HalluShift++ baseline follows the intuition of Nath et al. (2025): hallucinated content can be detected from internal model signals, including uncertainty, representation shifts, and attention behavior. Given an image, a question, and a generated answer, we perform a teacher-forced forward pass through an evaluating VLM. This evaluation model can either be the same LVLM that produced the answer, corresponding to a model-dependent self-probing setup, or a separate LVLM used to evaluate outputs generated by other models. In all cases, the original generation process remains unchanged; the forward pass is used only post hoc to extract token-level features.

For each response token, we extract uncertainty-, representation-, and attention-based features. The uncertainty features include the maximum softmax probability, its reciprocal as a confidence proxy, and the probability assigned to the observed token under teacher forcing. From the latter, we compute token-level negative log-likelihood and perplexity. The representation features include layer-wise hidden-state norms and layer-consistency scores, computed as the normalized cosine similarity between an early and a late decoder layer, together with their complements. The attention features include layer-wise attention means and entropies, as well as attention-concentration statistics based on the mean and standard deviation of Gini coefficients over the last attention layers. We additionally include sequence-level confidence and token-pattern features. These comprise the mean and standard deviation of the inverse maximum probability, the confidence trend across the answer, the mean confidence, the fraction of low-confidence tokens, repetition ratios, and normalized lexical diversity. Since our prediction target is token-level, sequence-level features are attached to every token of the corresponding answer.

To obtain token-level supervision, we align each decoded token with its character span in the generated answer. Tokens whose character span overlaps

Metric	Generating LVLM					Language			
	gemma3	internvl	llava	minicpm	qwenvl	EN	FR	IT	ZH
Presence	0.50 ± 0.02	0.49 ± 0.02	0.43 ± 0.01	0.47 ± 0.01	0.30 ± 0.01	0.47 ± 0.02	0.32 ± 0.02	0.46 ± 0.01	0.50 ± 0.02
Label (longest/all)	0.44 ± 0.01	0.42 ± 0.01	0.34 ± 0.01	0.39 ± 0.01	0.26 ± 0.01	0.37 ± 0.01	0.31 ± 0.01	0.39 ± 0.01	0.43 ± 0.02
Label (first/all)	0.46 ± 0.01	0.43 ± 0.01	0.38 ± 0.02	0.41 ± 0.01	0.28 ± 0.02	0.40 ± 0.01	0.33 ± 0.01	0.41 ± 0.01	0.44 ± 0.01
Label (longest/hallu.)	0.49 ± 0.03	0.47 ± 0.02	0.39 ± 0.02	0.42 ± 0.01	0.37 ± 0.01	0.38 ± 0.01	0.42 ± 0.01	0.42 ± 0.02	0.49 ± 0.02
Label (first/hallu.)	0.54 ± 0.03	0.50 ± 0.02	0.45 ± 0.02	0.48 ± 0.01	0.42 ± 0.03	0.46 ± 0.02	0.44 ± 0.02	0.47 ± 0.01	0.54 ± 0.01
Span (ignoring labels)	0.44 ± 0.02	0.45 ± 0.01	0.39 ± 0.01	0.38 ± 0.01	0.32 ± 0.01	0.35 ± 0.01	0.39 ± 0.01	0.43 ± 0.01	0.43 ± 0.01
Span (Requiring label match)	0.32 ± 0.01	0.33 ± 0.01	0.29 ± 0.01	0.29 ± 0.01	0.24 ± 0.01	0.25 ± 0.01	0.29 ± 0.01	0.31 ± 0.01	0.32 ± 0.01

Table 10: Annotator agreement grouped by generating LVLM and output language on the random and MAP subsets. See Table 9 for an overview of the metrics. IAA scores are reported as mean $\pm$ 95% bootstrap confidence interval.

		acc	prec	rec	F1
overall	**human**	0.664	0.576	0.508	0.499
	gemma3	0.543	0.505	0.444	0.442
	internvl	0.532	0.460	0.383	0.401
	llava	0.574	0.465	0.490	0.455
	minicpm	0.485	0.492	0.393	0.410
	qwenvl	0.459	0.503	0.412	0.427
English	**human**	0.695	0.583	0.544	0.511
	gemma3	0.522	0.686	0.397	0.391
	internvl	0.516	0.463	0.303	0.328
	llava	0.568	0.591	0.544	0.516
	minicpm	0.449	0.458	0.344	0.345
	qwenvl	0.402	0.405	0.322	0.307
French	**human**	0.678	0.577	0.518	0.509
	gemma3	0.539	0.442	0.533	0.450
	internvl	0.463	0.325	0.359	0.321
	llava	0.538	0.441	0.546	0.449
	minicpm	0.481	0.468	0.415	0.421
	qwenvl	0.427	0.513	0.481	0.458
Italian	**human**	0.650	0.582	0.487	0.482
	gemma3	0.539	0.505	0.475	0.452
	internvl	0.618	0.407	0.380	0.383
	llava	0.676	0.554	0.516	0.516
	minicpm	0.519	0.609	0.413	0.437
	qwenvl	0.433	0.450	0.523	0.458
Chinese	**human**	0.635	0.563	0.495	0.491
	gemma3	0.566	0.520	0.518	0.506
	internvl	0.532	0.597	0.458	0.505
	llava	0.508	0.365	0.449	0.390
	minicpm	0.493	0.422	0.386	0.393
	qwenvl	0.594	0.572	0.372	0.425

Table 11: Gemma3 classification full results.

an annotated hallucination span are labeled as hallucinated; tokens with no overlap are labeled as non-hallucinated. In the multiclass setting, hallucinated tokens are assigned the corresponding hallucination type. We then train a lightweight multilayer perceptron over the resulting token-level feature vectors, using the same classifier configuration as Nath et al. (2025). Our adaptation differs from the original HalluShift++ formulation in one main respect: while HalluShift++ decomposes generated descriptions into semantic chunks, such as object, attribute, and relation units, our annotations are

Strategy	acc	prec	rec	F1
random	0.659	0.180	0.672	0.284
MAP	0.633	0.268	0.700	0.387
human-written	0.565	0.263	0.846	0.401

(a) HalluShift++ performance (accuracy, recall, precision, F1) across strategies.

	random	MAP	all
gemma3	0.257	0.418	0.324
internvl	0.280	0.382	0.336
llava	0.398	0.402	0.401
minicpm	0.276	0.345	0.312
qwenvl	0.248	0.391	0.327

(b) HalluShift++ F1 per generating LVLM.

Table 12: HalluShift++ performance (binary label-classification setup).

span-based. We therefore align internal signals directly to decoded tokens and train a token-level classifier.

Performance in binary setting. In Table 12, we provide an overview of performances for the HalluShift++ method based on span detection, framing the problem as a binary token-labeling task instead (i.e., F vs. all). Similar conclusions emerge: a very low precision severely curtails the performances of these models. In Table 12b, we see a significant downgrade in performance for randomly selected items.

C.4 Span assessment with EUQ

Another baseline we consider is based on Evidential Uncertainty Quantification (EUQ; Huang et al., 2026). It provides a training-free framework that explicitly decomposes token-level epistemic uncertainty into different types of misbehaviors. The framework applies basic belief assignment (Dempster, 1967) to the token-level pre-logits feature obtained from the output head of LVLM to compute evidence weights. These weights are then decomposed into positive and negative components,

		acc	prec	rec	F1
strategy	**random**	0.794	0.196	0.198	0.197
	MAP	0.733	0.269	0.167	0.206
	human-written	0.719	0.267	0.215	0.238
random	**gemma3**	0.817	0.179	0.230	0.201
	internvl	0.804	0.201	0.219	0.210
	llava	0.708	0.331	0.177	0.230
	minicpm	0.799	0.191	0.213	0.201
	qwenvl	0.809	0.156	0.188	0.170
MAP	**gemma3**	0.753	0.323	0.173	0.226
	internvl	0.757	0.311	0.187	0.233
	llava	0.686	0.320	0.146	0.201
	minicpm	0.764	0.274	0.224	0.246
	qwenvl	0.735	0.214	0.152	0.178
all	**gemma3**	0.794	0.226	0.200	0.212
	internvl	0.780	0.256	0.198	0.223
	llava	0.694	0.324	0.157	0.211
	minicpm	0.782	0.232	0.219	0.226
	qwenvl	0.774	0.185	0.166	0.175

Table 13: Span detection with EUQ, token-level classification results.

which are fused to estimate the final uncertainties that can detect conflict and ignorance. Conflict refers to error caused by the difference in the image and text modality; whereas, ignorance refers to error caused by missing information. Using the optimal threshold obtained from a calibration set for conflict and ignorance, we filter spans for containing hallucinations. Results are shown in Table 13. The high label imbalance between hallucinated and non-hallucinated tokens leads to poor performances out-of-the-box, lower than what we observed with HalluShift++ in Table 12. For accuracy, recall, and precision, we can always find specific LVLM-based setups that either under- or over-perform with respect to the human-written samples subset, although the human-written data does yield the highest F1 score.

D Disclaimers

Use of existing artifacts. We rely on preexisting models and datasets to construct our own dataset; the use is compatible with explicit licenses.

Use of AI tools. AI technology was used to jumpstart coding. Authors take full responsibility for the contents of this paper.

Hardware and compute. Our experiments were run on a variety of compute clusters, including V100, A100, RTX A6000, MI250x, H1000.

LVLM answer generation for all models except Gemma3 was carried out on a node equipped with two NVIDIA RTX A6000 GPUs. Generation required approximately 4.5 seconds per sample across almost all models, whereas Qwen3-VL required approximately 9 seconds per sample on average.

LVLM answer generation for Gemma3 was carried out on a node equipped with one AMD MI250x GPU. Generation required approximately 6 seconds per sample on average.