

Sparse-Positive Hallucination Evaluation Entangles Detection and Prevalence

Srinivas Harish

New York University

srinivas.harish@nyu.edu

Abstract

Span-level hallucination metrics aggregate outputs whose references contain hallucinated spans with outputs whose references are empty. Under scoring rules that award a perfect match to an empty prediction against empty gold, the aggregate entangles detection performance with empty-gold prevalence. We write the exact decomposition as $s = ea + (1 - e)d$, where e is empty-gold prevalence, a is performance on empty-gold items, and d is performance on non-empty-gold items. Under the benchmark’s metric conventions, an empty submission identifies e from leaderboard values without access to closed labels. Applied to a multilingual vision-language hallucination benchmark, this recovers rates from 0.1549 to 0.2554. Assuming perfect empty-gold performance gives a maximum contribution equal to 47.3–64.8% of the mean team score and a conservative lower bound on d . Under this bound, the highest-ranked language moves from best for 17 of 18 teams to best for none and falls from rank 1 to rank 3. Released annotations show that union-over-annotators gold retains 4.09 times as many characters as majority-agreement gold, while 71.7% of annotated spans were marked by one of three annotators. Cross-split comparisons therefore require reporting both empty-gold prevalence and system performance on empty-gold items alongside aggregate scores.

1 Introduction

Span-level hallucination detection asks a system to mark which characters of a model’s output are unsupported. Many items contain nothing to mark, and a scorer has to decide what a correct empty prediction is worth. SHROOM-Visions (Vázquez et al., 2026; Mickus et al., 2026) decides that it is worth a perfect score: all three of its metrics return 1.0 when reference and prediction are both empty. Every published number is therefore a sum of two things, credit for abstaining correctly, and

performance on items that actually contain a hallucination.

The two terms can be separated without access to the closed test annotation, because the first one leaves a signature. This paper separates them, and the separation changes what the leaderboard says.

The empty-gold rate is recoverable from the leaderboard. An empty submission scores 1.0 on every empty-gold item and 0.0 on every other, so its three metric values are identical and all equal the empty-gold rate. Two rows carry that signature in all four languages, putting the test split’s rate at 0.1549–0.2554 depending on language.

The maximum empty-gold contribution is large, and it does not cancel. Assuming perfect empty-gold performance, it is 47.3–64.8% of the mean team score. Subtracting this maximum yields a lower bound on non-empty-gold performance and reverses the cross-language ordering: the language strongest for 17 of 18 teams on the label-aware metric is strongest for 0 of 18 under the bound; on IoU, 18 of 18 becomes 1; and it falls from rank 1 to rank 3 of four.

Gold construction controls which items count as empty. The scorer takes the union of every annotator’s marks with no agreement threshold, so an item is empty only when no annotator marked anything at all. 71.7% of the spans that keep an item non-empty were marked by one annotator of three. This mechanism can affect the rate, but the available data do not establish why it differs across languages.

2 The Metrics and the Empty-Submission Score

We use the organisers’ scoring functions unmodified, as shipped in the participant kit. `score_iou` forms $|R \cap P|/|R \cup P|$ over character index sets and returns 1.0 when both are empty; `score_cor` takes Spearman correlation between per-character

annotator-probability vectors and returns 1.0 when both are constant and zero; `score_cor_lbl` averages that correlation over categories and returns 1.0 when neither side carries labels.

The signature. Let e be the fraction of items whose gold is empty. An empty submission scores 1.0 on exactly those items and 0.0 on every other, under all three metrics. Its three values are therefore identical and all equal e . We verified this identity on the released training split, where e is directly observable: the empty submission scores 0.253–0.365 and the three metrics agree to the last digit in every language.

No other submission has that property in general, and on the closed leaderboard exactly two rows do, `baseline` and `WINNAR749`, with the same value in each language. Reading them off gives the test split’s empty-gold rate: 0.1549 (EN), 0.1679 (FR), 0.1635 (IT), 0.2554 (ZH).

A trap for anyone reproducing this. The `scorer.py` in the task’s public repository is the previous edition’s. The 2026 scorer ships only in the participant kit, and the two differ in the place that matters: the earlier one passes the reference through a `prob > 0.5` filter before computing IoU, the current one does not.

3 Structural Origin of the Coupling

Nothing above depends on the overlap function. Let M be any span-level metric that scores a prediction P against a reference R . Write e for the fraction of items whose reference is empty, a for a system’s mean score on those items, and d for its mean score on the remaining items. Every system’s aggregate score is exactly

$$s = ea + (1 - e)d. \quad (1)$$

For metrics bounded by one, $a \leq 1$ implies $d \geq (s - e)/(1 - e)$. Thus the empty-submission score recovers e , while subtracting e from another system’s score gives a conservative lower bound on its non-empty-gold performance rather than an exact decomposition of that system’s behavior.

The three SHROOM-Visions metrics are structurally unrelated: `score_iou` is a set overlap, `score_cor` is a rank correlation over per-character probabilities, and `score_cor_lbl` averages that correlation over label categories. Yet an empty submission scores the same e under each.

This shared empty-case behavior, rather than any particular overlap function, makes e recoverable.

The convention is not obviously wrong. Abstaining correctly is a real capability, and a scorer that returned 0 for an empty match would punish it. What the sum does is make one number answer two questions, and the fix is to report the two terms rather than to change the convention.

4 Bounding Non-Empty-Gold Performance

A published score decomposes as $s = ea + (1 - e)d$. Because $a \leq 1$, the quantity $\hat{d} = (s - e)/(1 - e)$ is a lower bound on performance d over items that contain a hallucination. It is computable for every team from published numbers alone and is exact when $a = 1$.

Table 1 and Figure 1 apply it to the 18 teams carrying a score on all three metrics in all four languages. On `score_cor_lbl`, the label-aware metric, 47.3–64.8% of the mean team score is abstention credit, and the share tracks the empty-gold rate: it is largest in the language whose annotators most often marked nothing. That language is the strongest for 17 of 18 teams as published and 0 of 18 after correction, falling from rank 1 to rank 3. The same holds on the other two metrics: IoU moves 18→1 teams and rank 1→3, the correlation metric 13→2 and rank 1→4.

The bound transformation is monotone within a language, so it does not reorder teams competing on the same split; every within-language placement is unchanged. What it reorders is the comparison *across* languages, which is exactly the comparison a reader makes when the leaderboard shows four tables side by side and the same column is the highest for almost every team.

5 Cross-Language Variation in the Empty-Gold Rate

An item’s gold is empty only if no annotator marked anything, because `score_iou` builds its reference from every annotated span regardless of how many annotators contributed it. The released training split, the only one whose annotation is public, is 15 102 items across the four languages, 10 807 of them carrying at least one span. It carries 41 998 annotated spans in total, of which 30 120 (71.7%) have an empirical probability of one third: one annotator of three marked that span, and the other two did not (Table 2). A single annotator

Lang	recovered	as published		lower bound		maximum
	e	mean	rank	mean	rank	credit (% of mean)
EN	0.1549	0.3093	4	0.1827	4	50.1
FR	0.1679	0.3428	3	0.2102	2	49.0
IT	0.1635	0.3455	2	0.2175	1	47.3
ZH	0.2554	0.3943	1	0.1865	3	64.8

Table 1: **Bounding non-empty-gold performance on the closed leaderboard.** e is recovered from the empty-submission signature. Means are over the 18 teams carrying all scores; the two signature rows are excluded. Higher is better. The quantity $(s - e)/(1 - e)$ is a lower bound on non-empty-gold performance, exact when a system scores 1 on every empty-gold item. The final column is the maximum empty-gold contribution as a share of the mean.

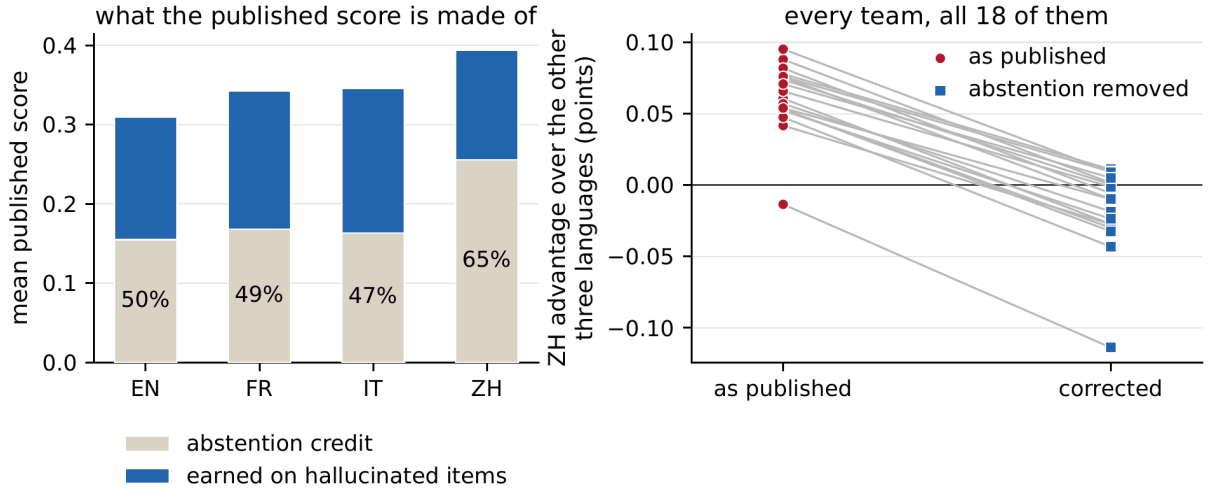


Figure 1: **Maximum empty-gold contribution and its effect on language comparison.** Left: each language’s mean published `score_cor_lbl` under the conservative assumption $a = 1$; the percentage is the maximum empty-gold share. Right: each team’s ZH contrast as published and after subtracting that maximum. The plotted “detection only” values are therefore lower bounds, exact when empty-gold performance is perfect. Exact means and ranks appear in Table 1.

marking a single character is enough to move an item out of the empty-gold class and remove a system’s free point.

This is a change from the previous edition, and a change away from a documented rule: MuSHROOM binarized the gold at majority agreement before computing IoU and gave the equation for it (Vázquez et al., 2025), while the SHROOM-Visions scorer applies no threshold. The same annotation yields a gold $4.09\times$ larger. The majority-agreed core is 24.5% of the union gold, character-weighted, with a 95% interval of 23.4–25.6 from 5,000 resamples of items, recomputing the ratio of summed characters on each; averaging per-item ratios instead gives 31.4%. Requiring annotators to agree on the category as well gives 18.0%, and Table 3 gives the retained share at every cut a threshold rule could take. A detector reproducing the annotator majority exactly recovers 0.297–0.337

IoU on items that contain a hallucination.

Which aggregation an evaluation should target is an open question in its own literature. Plank (2022) argues against collapsing annotator variation into a single ground truth; Baan et al. (2022) shows that scoring against a human majority is the wrong instrument where annotators genuinely disagree; Kurniawan et al. (2024) compares majority-voted and best-matched span golds for exactly this kind of task. None of the three scores the union, which is what this metric does, and none measures what the choice costs on a leaderboard, which is what Section 4 does.

6 Discussion

Report the empty-gold rate beside the score. It is one number per split, the organisers already have it, and it exposes a major source of cross-split variation. With it, every reader can compute the lower

Lang	spans	spans by agreement, n (%)			4-ann. spans
		1 of 3	2 of 3	3 of 3	
EN	12894	9171 (71.1)	2661 (20.6)	963 (7.5)	99
FR	12216	8846 (72.4)	2628 (21.5)	692 (5.7)	50
IT	9359	6777 (72.4)	1756 (18.8)	780 (8.3)	46
ZH	7529	5326 (70.7)	1541 (20.5)	652 (8.7)	10

Table 2: **Every annotated span, by how many annotators contributed it.** The last column counts spans whose probability is a multiple of one quarter, i.e. items that received four annotators rather than three. They carry 0.56% of annotated characters, which account for all of the movement between the > 0 and $> 1/3$ rows of Table 3.

Agreement cut	gold retained (%)	regime
> 0	100.0	union
$> 1/4$	99.6	union
$> 1/3$	99.6	union
$> 1/2$	24.5	majority
$> 2/3$	6.1	unanimous
> 0.99	6.0	unanimous

Table 3: **Share of the union gold retained at each agreement cut,** character-weighted over annotated items. Three annotators admit only the probabilities $1/3$, $2/3$ and 1 , so the rule selects among three discrete golds rather than moving along a continuum; the residual movement between those points is the 0.56% of characters from items carrying four annotators. The current scorer applies no cut and sits in the first row.

bound $(s - e)/(1 - e)$; reporting empty-gold performance a as well would identify d exactly.

A “systems do best in language L ” reading needs the correction. That claim is what the published tables show and is not what the underlying detection shows. It is a property of how often L ’s annotators marked nothing, amplified by a scoring convention that pays full credit for agreeing with them, and it survives on all three metrics.

Cross-edition IoU is not a comparison. MuSHROOM binarized at majority and stated the rule; SHROOM-Visions does not binarize. Numbers computed against golds differing by a factor of 4.09 on the same annotation do not measure progress against each other.

7 Conclusion

All three SHROOM-Visions metrics give a perfect score to an empty prediction on empty gold. The empty-gold rate is recoverable from the leaderboard because an empty submission scores identically on all three metrics and two rows carry that

signature. Assuming perfect empty-gold performance yields a maximum contribution of 0.1549–0.2554 depending on language, or 47.3–64.8% of the mean team score. Subtracting that maximum gives a conservative lower bound that takes the apparently strongest language from best for 17 of 18 teams to 0 of 18 and from rank 1 to rank 3. The union-over-annotators construction determines which items are empty, but the available data do not identify the cause of cross-language variation. Reporting both the empty-gold rate and system performance on empty-gold items would make the two components identifiable.

Limitations

The empty-gold rate is recovered from two leaderboard rows whose three metrics coincide, which is the algebraic signature of an empty submission but is not proof that those rows are empty submissions; coincidence could misidentify them. For an arbitrary system, the exact decomposition is $s = ea + (1 - e)d$. Because empty-gold performance a is not published, $(s - e)/(1 - e)$ is only a lower bound on d , exact when $a = 1$. The gold-composition figures are computed on the released training split, whose empty-gold rate is higher than the recovered test rate, so they characterize the annotation protocol rather than establish the cause of test-set language differences. They are character-weighted, unlike a per-item mean. The annotator count is inferred from observed probability values rather than documented: almost every item carries three, and a small set carries four, contributing 0.56% of annotated characters.

References

- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *EMNLP*.
- Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2024. To aggregate or not to aggregate. that is the question: A case study on annotation subjectivity in span prediction. In *WASSA*.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *EMNLP*.
- Raúl Vázquez, Aman Sinha, Chuyuan Li, et al. 2026. Overview of SHROOM-Visions 2026: A shared task on hallucination detection in large vision-language models. *arXiv:2608.25662*.

Timothee Mickus, Claudio Savelli, Eduardo Calò, et al. 2026. Can humans dream of electric sheep? Human-written samples for fine-grained vision-and-language hallucination benchmarking. *arXiv:2608.01021*.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, et al. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *SemEval*.

A Reproducibility

All figures come from the released training split and the organisers’ own scoring functions as shipped in the participant kit, imported without modification. The measurement scripts and the derived JSON are in the accompanying archives; `claims.json` maps each numbered claim to the script that regenerates it. Note that the `scorer.py` in the task’s public GitHub repository is the previous edition’s and will not reproduce these numbers; the participant-kit copy is the one used here.