

TÜRKSAT at SHROOM-Visions: Crowd-Emulating Judge Ensembles for Multilingual Hallucination Span Detection in Vision-Language Models

Alper Karamanlioğlu
TÜRKSAT
Ankara, Türkiye
alper.karamanlioglu@
turksat.com.tr

Gökhan Yurtalan
TÜRKSAT
Ankara, Türkiye
gokhan.yurtalan@
turksat.com.tr

Berkan Demirel
TÜRKSAT
Ankara, Türkiye
berkan.demirel@
turksat.com.tr

Abstract

We describe our submission to SHROOM-Visions 2026, the shared task on generator-agnostic, character-level hallucination span detection and typing in vision-language model outputs across English, French, Italian, and Chinese. Rather than training a detector, we use a *crowd-emulating judge ensemble*: one instruction-tuned multimodal model, sampled at several temperatures and with several prompt variants, independently re-inspects each image, prompt, and response and proposes typed hallucination spans. Character-level vote fractions form soft probabilities that emulate the multi-annotator gold labels, while a cross-validated threshold per language produces hard spans. The black-box system requires no parameter updates and relies on guideline-derived prompt rules and language-specific few-shot examples. On the official closed test set, it ranks first for all English metrics and for Italian IoU, and second for every remaining language-metric combination. Official scores range from 0.4336 to 0.4845 for Cor+Lbl, 0.5474 to 0.5814 for Cor, and 0.4865 to 0.5240 for IoU. On held-out training folds, IoU is 0.32 to 0.39 on non-empty items. Ablations identify prompt design and few-shot conditioning as decisive, while an added adjudication stage does not improve performance. Vote fractions track annotator agreement but remain globally over-confident (ECE 0.08 to 0.12), motivating per-language thresholding.

1 Introduction

Vision-language models (VLMs) fluently describe images but routinely assert content the image does not support: inventing objects, miscounting, misreading text, or misattributing attributes. SHROOM-Visions 2026 (Vázquez et al., 2026) is the first multimodal edition of the SHROOM shared task series. Given an image, a prompt, and a model response, participants must (a) mark the

character spans of the response that are hallucinated and (b) assign each span one of five types: INVENTION, MISCHARACTERIZATION, OCR, MISCOUNTING, or OTHER. Systems are ranked *per language*. The task defines a span-identification metric (character intersection-over-union, IoU) and a calibration metric (Spearman ρ between predicted probabilities and the empirical multi-annotator gold soft labels); the official scoreboard we received reports two correlation variants, ρ and a label-conditioned ρ_{lbl} , as the ranked scores (Section 6).

A question posed by the organisers is whether systems designed for the text-only Mu-SHROOM task (Vázquez et al., 2025) transfer to the multimodal setting. Our training-fold results suggest the *crowd-emulation* strategy that characterised the strongest Mu-SHROOM entries, especially MSA (Hikal et al., 2025) and UCSC (Huang et al., 2025), adapts naturally to images: a panel of judges independently annotate each response, and their agreement becomes a calibrated soft label. The official closed-test Cor, Cor+Lbl, and IoU results confirm this transfer across all four languages (Section 6).

We make no use of internal model states (unlike the provided HalluShift++ baseline (Nath et al., 2025)) and train no parameters; the system is a prompted, generator-agnostic ensemble. Our contributions are: (1) a character-level vote-aggregation scheme whose output emulates gold multi-annotator soft labels; (2) a prompt design, distilled from the annotation guidelines and a manual study of the gold data, that drives most of our performance; (3) an empirical finding that an LLM *adjudication* stage does not help in our multimodal implementation, so that plain majority voting is both cheaper and at least as strong; and (4) a language-specific few-shot and thresholding recipe with an honest decomposition of where the score comes from.

2 Background and Related Work

Hallucination span detection. Mu-SHROOM (Vazquez et al., 2025) framed hallucination detection as character-level span identification with multi-annotator soft labels. The overview reports that retrieval-augmented and fine-tuned systems tended to outperform naive prompting on IoU, and that crowd-emulation ensembles of LLM “annotators” were used by several top systems, including UCSC (Huang et al., 2025) and MSA (Hikal et al., 2025). Top IoU scores there ranged roughly 0.53 to 0.78 across languages, above the range we obtain here, while human inter-annotator agreement was 0.31 to 0.51 IoU, on the text-only data. We reference these numbers as context, not as a directly comparable ceiling: they were measured on a different (text-only) task and, for the human band, are human-vs-human rather than system-vs-gold.

Relation to self-consistency and LLM-as-judge. Because our “panel” is one model resampled across temperatures and prompt variants rather than distinct models, it is closely related to self-consistency sampling (Wang et al., 2023) and to the broader use of LLMs as evaluators (Zheng et al., 2023). Our contribution is not the ensemble mechanism itself but (i) casting the character-level vote fraction as an emulator of multi-annotator soft labels, (ii) the calibration analysis of that estimator, and (iii) the negative adjudication finding.

Hallucination in VLMs. A growing literature evaluates or localises VLM hallucinations, from object-hallucination benchmarks (POPE (Li et al., 2023), CHAIR (Rohrbach et al., 2018)) to internal-signal methods (HalluShift++ (Nath et al., 2025)) and token-level supervision (HalLoc (Park et al., 2025)). The internal-signal methods assume white-box access to the generating model. SHROOM-Visions is generator-agnostic: the VLMs that produced the responses are not disclosed, so internal-state methods are only approximately applicable, motivating our black-box, judge-based design.

3 Task and Data

The shared task is built on the SHEEP dataset (Mickus et al., 2026), which pairs human-written and model-generated vision-and-language outputs with fine-grained, span-level hallucination annotations across four languages. The organisers release $\sim 15,100$ labelled training items ($\sim 3,800$ per language) and a closed test set of $\sim 4,900$ items, with

source images. Each item has an image reference, a prompt, a model response, and (for training) a list of typed, character-indexed spans with a per-span probability. Probabilities take annotator-fraction values $(\frac{1}{3}, \frac{2}{3}, 1)$, reflecting how many of three annotators marked a character. Roughly one quarter of responses carry *no* hallucination (empty label list); the official scorer awards $\text{IoU} = 1$ when both reference and prediction are empty, so correctly abstaining on clean responses earns full credit. As we show in Section 6, this makes the aggregate IoU substantially easier than the underlying detection problem, and we report a decomposition accordingly.

Observations that shaped our design. A manual study of the gold data revealed consistent annotator behaviour we later encoded as prompt rules: (i) annotators mark only claims that *contradict or cannot be verified against the image*, generally leaving unmarked world-knowledge errors, fabricated off-image content, and degenerate text; (ii) when the direct answer to the prompt is wrong, the answer-bearing tokens are marked with high agreement; (iii) spans are typically short (a few content words). We also noted that image filenames often encode the depicted object’s unusual property (e.g. `Three-legged_cat`); the rules permit us to pass the filename as a hint. A small number of Italian test items have no associated image and necessarily receive empty predictions.

4 System Description

4.1 Judge panel

For each item we query a panel of K judges. Each receives the image, the prompt, the response, an instruction block, and a few few-shot examples, and returns a JSON list of spans, each with a verbatim text substring, a locating context string, and a type label (schema in Section 4.3). We obtain the diversity that makes voting useful from two axes: sampling temperature (we use 0.2, 0.6, 1.0) and, for some languages, prompt variants. The “panel” is thus a *single* instruction-tuned multimodal model queried repeatedly, not a mixture of distinct models; we found this more stable than mixing in weaker models. We aggregate $K = 6$ votes per item for EN/FR and $K = 9$ for IT/ZH, obtained by sampling the prompt variants of Section 4.3, namely the fixed-format **v4** and occurrence-format **v2o** configurations, plus a k -NN-retrieval configuration for

IT/ZH, across the three temperatures (all three variants are described in Section 4.3).

4.2 Character-level vote aggregation

Each judge’s spans are mapped to character offsets by locating the returned substring in the response, disambiguated by the context string; spans whose text cannot be located are dropped. For a response of length L we accumulate, per character i , the number of judges v_i covering i (a judge covers i at most once, even with overlapping spans). The soft probability is $p_i = v_i/K$. Because the gold probabilities are themselves annotator fractions, a K -judge vote fraction is a natural emulator of the multi-annotator soft label that ρ rewards. To produce hard spans for IoU we threshold: maximal runs with $p_i \geq \tau$ become predicted spans, each assigned the majority type among its covering judges (typing is decided *per character*, then runs are formed). The single hyperparameter τ is tuned per language (Section 5). We tested boundary post-processing (gap-filling, whitespace/punctuation trimming) and snapping p_i to the nearest annotator-fraction bucket; none gave a reliable gain (Section 7), so the deployed system uses none.

4.3 Prompt design and language recipes

We reproduce the deployed prompt so the paper is self-contained. The judge receives the image, the user prompt, and the model response, followed by the instruction block below.

Instruction block (shared across variants).

1. Look at the image carefully first. Count objects yourself, read any text character by character, and check colours/attributes/positions before judging.
2. Mark a span *only* if it describes the image incorrectly: it contradicts what is visible, or claims to see something that is not there.
3. Do *not* mark: general world knowledge (even if wrong or unverifiable, e.g. “likely a Manx”); fabricated content not presented as a description of the image (invented schedules, general information, disclaimers); degenerate/nonsense/repetitive text, empty answers, refusals; discourse fillers (“Based on the image”, “Yes,”).
4. Most important: if the direct answer to the question is wrong (wrong number, colour, lo-

cation, or a nonexistent object), you **MUST** mark the answer-bearing words themselves (the number, colour, side, or object name), every time they appear.

5. Also mark subtle wrong visual details inside otherwise correct sentences (wrong side/position, wrong count of a secondary object).
6. Keep spans minimal: the specific wrong words, not whole sentences.
7. Classify each span as INVENTION, MISCHARACTERIZATION, OCR, MISCOUNTING, or OTHER.
8. If nothing qualifies, return an empty JSON array.

Filename hint (variant-dependent). Some variants additionally pass the image filename, noting that filenames often name the unusual property of the object (e.g. Three-legged_cat), to be used only as a hint and verified against the image.

Prompt variants. Three configurations provide the voting diversity, differing only in output format and few-shot set. **v4** requests, per span, a verbatim text, a longer context substring containing it exactly once, a label, and a confidence (*high/low*), and includes the filename hint. **v2o** instead uses text, a 1-based occurrence index, and a label, with no filename hint. **kNN** (used for IT/ZH) uses the v4 format but replaces the fixed few-shot examples with the $k=3$ nearest training items (char n -gram TF-IDF over prompts, $n \in \{2, 3, 4\}$). EN/FR aggregate v4 and v2o ($K=6$); IT/ZH aggregate v2o and kNN ($K=9$). Low-confidence spans are down-weighted to 0.4 in aggregation.

Output schema (v4).

```
[{"text": "<verbatim substring>",
 "context": "<longer verbatim substring
  containing text exactly once>",
 "label": "<invention|mischaracterization|
  OCR|miscounting|other>",
 "confidence": "<high|low>"}]
```

Few-shot examples. Five hand-written examples are shown to the judge (image described in brackets); for Chinese they are localised. We give them verbatim:

A [marina with 5 boats]. Prompt: “How many boats are there?” Response: “...at least **6 boats**. ...Another partially visible boat on the right. ...at least 6.” → mark “6” (MISCOUNTING), “Another partially visible boat on the right” (INVENTION), and the second “6” (MISCOUNTING).

B [a wristwatch; response is degenerate text]. Response: “://www.google.com/2024/03/21/02/...” → mark nothing (nonsense text is not a wrong description of the image).

C [concert hall exterior; response invents a schedule]. Response: “As of today, November 21, 2023, upcoming events: ...” → mark nothing (fabricated general information, not a wrong description of the image).

D [a bird standing right of centre]. Response: “...positioned ... just to the left of the center.” → mark “to the left of the center” (MISCHARACTERIZATION).

E [a purple cauliflower; asked if white]. Response: “No ... it is a vibrant **purple**.” → mark nothing (the trick-question answer is correct).

5 Experimental Setup

All judges are a single flash-tier multimodal model, Gemini 3.5 Flash (Google DeepMind, 2026), queried through a public API at temperatures $\{0.2, 0.6, 1.0\}$, with a capped thinking budget of 1024 tokens and a 3072-token output limit for cost and latency control; Gemini 2.5 Pro (Comanici et al., 2025) at temperature 0.2 serves as the adjudicator in the ablation of Section 7. The judge diversity aggregated into each vote comes from these three temperatures crossed with the prompt variants of Section 4.3 ($K=6$ for EN/FR, $K=9$ for IT/ZH). Each variant optionally emits a per-span confidence flag; “low”-confidence spans enter aggregation with weight 0.4 rather than 1.0, softening borderline marks. The pipeline is fully resumable, and producing all test predictions required roughly 29k API calls in total. We evaluate on a fixed random sample of ~ 400 labelled training items per language (a cost-driven subset of the $\sim 3,800$ available) via 5-fold cross-validation: for each fold we select τ on the other four folds and score on the held-out fold, reporting the mean and standard deviation across folds. **Scope of tuning.** Only τ is cross-validated. The prompt rules, filename hint, the fixed-vs- k NN choice, K , and the temperature set were all selected on training data and may therefore be optimistic; we did not run nested cross-validation over these. For k -NN languages, the retrieval index *excludes all evaluation items* (the judge for a held-out item can never retrieve that item), but same-image or near-duplicate prompts elsewhere in the training pool could still leak mild signal; the reported IT/ZH numbers should be read with this caveat.

6 Results

Table 1 reports the official TÜRKSAT leaderboard results. On cross-validated development folds, ag-

Lang	Cor+Lbl (rank)	Cor (rank)	IoU (rank)
EN	0.4336 (1)	0.5534 (1)	0.4865 (1)
FR	0.4511 (2)	0.5474 (2)	0.5000 (2)
IT	0.4437 (2)	0.5532 (2)	0.4902 (1)
ZH	0.4845 (2)	0.5814 (2)	0.5240 (2)

Table 1: Official leaderboard results for TÜRKSAT. Parentheses give the rank within each language and metric.

gregate IoU is 0.44 to 0.49, but the $\sim 25\%$ empty-gold responses score 0.70 to 0.80 through correct abstention; on the actual non-empty detection problem, IoU is 0.32 to 0.39 versus 0 for the mark-none baseline. Correlation ρ is 0.49 to 0.54, while typed ρ_{lbl} is lower (0.41 to 0.45). Fold variance reaches ± 0.05 , so we avoid interpreting differences below 0.02.

Official test results. The published leaderboard reports Cor+Lbl, Cor, and IoU for each language. TÜRKSAT ranks **first in all three English metrics and in Italian IoU**, and **second in every remaining language-metric combination**. Table 1 gives only TÜRKSAT’s exact official values and ranks. Test Cor exceeds our development estimate in every language; the ~ 400 -item CV subset is noisier than the $\sim 4,900$ -item test set.

7 Analysis and Ablations

Prompt design is the dominant lever. Iterating the instruction block moved EN IoU from 0.385 (naive prompting) to 0.428 (guideline-derived rules and few-shot) and 0.436 (filename hint). Self-answering and longer prompts hurt every language, favouring concise, behaviour-matched instructions.

Does the ensemble help? Relative to one judge, the panel improves FR (0.461 $\rightarrow$ 0.487), IT (0.427 $\rightarrow$ 0.476) and ZH (0.467 $\rightarrow$ 0.482), but is neutral for EN (0.443 $\rightarrow$ 0.440). Its main benefit is therefore graded probabilities and stability, not universal IoU gains.

Adjudication does not help. Adding a separate Pro-tier adjudicator, Gemini 2.5 Pro, did not improve upon the Flash-based judge ensemble: MSA-style re-scoring (Hikal et al., 2025) lowered IoU in every mixture (e.g. 0.41 vs. 0.45 on EN) and roughly doubled cost, plausibly because its visual-grounding errors correlate with the panel’s.

Saturation and error sources. Post-processing changes IoU by at most 0.001. Of our 80 worst

Lang	gold $\frac{1}{3}$	gold $\frac{2}{3}$	gold 1	ECE
EN	0.33	0.69	0.90	0.084
FR	0.35	0.68	0.85	0.107
IT	0.41	0.74	0.97	0.121
ZH	0.45	0.79	0.95	0.088

Table 2: Calibration of the vote fraction p_i . Columns 2 to 4: mean p_i on characters whose gold annotator fraction is $\frac{1}{3}$, $\frac{2}{3}$, 1; the values are near-diagonal (good agreement-rank estimator). Last column: absolute ECE over all characters (over-confident; the panel over-marks), motivating the per-language τ .

items, $\sim 36\%$ reflect gold noise, $\sim 33\%$ precision loss and $\sim 26\%$ recall loss. Rare OCR/OTHER labels also contribute to the typed-correlation gap.

What the vote fraction measures. On gold-marked characters, p_i closely tracks annotator agreement: $\frac{1}{3}$ maps to 0.33 to 0.45, $\frac{2}{3}$ to 0.68 to 0.79, and 1 to 0.85 to 0.97 (Table 2). Globally, however, the panel over-marks (ECE 0.08 to 0.12). Thus p_i is a useful agreement-rank estimator but a poorly calibrated absolute probability, motivating language-specific τ (0.15 to 0.50).

8 Conclusion

Crowd emulation transfers effectively to multi-modal SHROOM-Visions. TÜRKSAT ranks first in all English metrics and Italian IoU, and second in every other language-metric combination. Official IoU ranges from 0.4865 to 0.5240. Prompt design and language-specific few-shot are decisive; adjudication adds cost without gains. An open-model replication is the main next step.

Limitations

Overall IoU is officially scored on the closed test set, but the empty versus non-empty decomposition and all ablations remain training-fold estimates. Only τ is cross-validated; other design choices were tuned on training data and may be optimistic. For k -NN languages, residual leakage from same-image or near-duplicate training prompts cannot be fully excluded. The system depends on a proprietary multimodal API, raising cost and reproducibility concerns; the filename hint may be an artifact whose test-set validity is unverified. Finally, a substantial share of residual error is intrinsic annotation disagreement that no system can resolve.

References

- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*. <https://arxiv.org/abs/2507.06261>.
- Google DeepMind. 2026. Gemini 3.5 Flash Model Card. <https://deepmind.google/models/model-cards/gemini-3-5-flash/>.
- Baraa Hikal, Ahmed Nasreldin, and Ali Hamdi. 2025. MSA at SemEval-2025 Task 3: High Quality Weak Labeling and LLM Ensemble Verification for Multilingual Hallucination Detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 989 to 995.
- Sicong Huang, Jincheng He, Shiyuan Huang, Karthik Raja Anandan, Arkajyoti Chakraborty, and Ian Lane. 2025. UCSC at SemEval-2025 Task 3: Context, Models and Prompt Optimization for Automated Hallucination Detection in LLM Output. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1981 to 1992.
- 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 to 305.
- Timothee Mickus, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Chuyuan Li, Aman Sinha, Lorenzo Vaiani, Jörg Tiedemann, and Raúl Vázquez. 2026. Can Humans Dream of Electric Sheep? Human-Written Samples for Fine-Grained Vision-and-Language Hallucination Benchmarking. *arXiv preprint arXiv:2608.01021*. <https://arxiv.org/abs/2608.01021>.
- 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*.
- 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 to 29903.
- 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 to 4045.
- Raúl Vázquez, 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: MUSHROOM, 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 to 2497.
- Raúl Vázquez, Aman Sinha, Chuyuan Li, Artem Shelmanov, Artem Vazhentsev, Claudio Savelli, Eduardo Calò, Emilio Raimond, Stella Frank, Hengyu Luo, Flavio Giobergia, Vincent Segonne, Lorenzo Viani, Jörg Tiedemann, and Timothee Mickus. 2026. Overview of SHROOM-Visions 2026: A Shared Task on Hallucination Detection in Large Vision-Language Models. *arXiv preprint arXiv:2608.25662*. <https://arxiv.org/abs/2608.25662>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*.