

Team hccl-jy at SHROOM-Visions: Confidence Signals from an Unmodified VLM Verifier

Junyoung Jeon Bongwon Suh

Department of Intelligence and Information

Seoul National University

{junn103.jeon, bongwon}@snu.ac.kr

Abstract

Vision-language models (VLMs) can describe details an image does not support. Detecting such hallucinations requires locating the unsupported text rather than flagging whole responses. The SHROOM-Visions shared task asks systems to localize these spans, classify their error type, and assign character-level probabilities. This paper describes the hccl-jy English-track submission. We ask whether confidence from a pretrained VLM verifier can support this task without fine-tuning, and whether candidate-level gains survive extraction and aggregation. The system runs an unmodified Qwen2.5-VL verifier over spaCy candidates with exact character offsets. An aggregation model fitted on held-out calibration data converts candidate judgments into output spans. We compare verbalized confidence, verdict-token probability, and five-draw consistency against annotator agreement, calibration error, and the official metrics. Verdict-token probability separates high-agreement candidates from low-agreement ones more reliably at 32B than at 7B. Averaging repeated judgments improves this ranking further but does not improve the official score. Error analysis shows that the initial score comes mostly from correct empty predictions and that error typing is weak for invented content. Revising only the aggregation over cached judgments raises Cor on development and test data. The findings support evaluating confidence discrimination, calibration, and end-to-end performance as separate targets.

1 Introduction

Hallucination detection covers several settings and units of analysis (Ji et al., 2023). Early visual work measured invented objects in image captions (Rohrbach et al., 2018). Later benchmarks examined object hallucination in VLMs (Dai et al., 2023; Li et al., 2023) and extended annotation to attributes and relations (Gunjal et al., 2024). SHROOM-Visions goes further by asking systems

to identify the exact unsupported characters, assign one of five error types, and predict a probability for every character.

The task belongs to the SHROOM series. The first shared task evaluated hallucination decisions for complete text responses (Mickus et al., 2024), and Mu-SHROOM introduced multilingual character spans and soft labels (Vázquez et al., 2025). SHROOM-Visions uses the SHEEP dataset and replaces textual grounding with images (Mickus et al., 2026). The task overview describes the data, metrics, and participating systems (Vázquez et al., 2026).

Related text methods evaluate atomic facts (Min et al., 2023), word-level hallucinations (Niu et al., 2024), or typed factual errors (Mishra et al., 2024). Visual methods also extract concepts and verify them against image evidence (Yin et al., 2024), while recent work probes internal VLM representations before generation (Kogilathota et al., 2026). Our system retains exact character offsets and studies uncertainty from a verifier after generation.

We ask whether confidence from a pretrained VLM verifier can support detailed hallucination detection without fine-tuning that verifier. We compare three confidence signals from the same verifier. The first is verbalized confidence, motivated by prior work on confidence elicitation (Lin et al., 2022; Xiong et al., 2024). The second is verdict token probability, related to work that uses probabilities such as $P(\text{True})$ to estimate answer correctness (Kadavath et al., 2022). The third is verdict consistency across repeated verifier calls, an analogue of hallucination detection based on repeated samples (Manakul et al., 2023). Holding the verifier fixed keeps its visual reasoning ability constant across comparisons.

Our pipeline combines candidate extraction, visual verification, and output aggregation. We compare verbalized confidence and verdict token probability across the 7B and 32B checkpoints, then



Figure 1: Sequence of experiments, read left to right and then along the second row. Orange boxes are VLM inference runs, blue boxes are offline analyses of cached judgments, and dark boxes are leaderboard submissions.

evaluate five-draw averaging and verdict consistency at 32B. We relate these signals to annotator agreement rather than only binary correctness. We also examine why better confidence for individual candidates does not necessarily improve the official score. Empty predictions on clean responses dominate the initial system, and the verifier’s error-type labels may limit Cor+Lbl. Revising the aggregation over cached judgments raises Cor on both development and test data without another VLM run. Figure 1 summarizes the sequence of runs and analyses.

2 Task and Data

SHROOM-Visions (Vázquez et al., 2026) uses the multilingual SHEEP dataset (Mickus et al., 2026). Each item contains an image, a prompt, and a response generated by a VLM. A submission gives

character offsets that include the start and exclude the end, a hallucination probability, and one label from *invention*, *mischaracterization*, *OCR*, *miscounting*, and *other*. The shared task evaluates character overlap (IoU), Spearman correlation between predicted and reference probabilities (Cor), and correlation computed separately by error label (Cor+Lbl).

We participate only in English. The labeled data contain 3,799 responses and 12,894 gold spans. We divide them by image into 2,627 calibration items and 1,172 development items. The two sets contain 1,125 and 482 images, respectively, and use seed 13. Responses that share an image cannot appear in different sets. Of all responses, 25.3% contain no hallucination. Most spans belong to the two less objective categories. Mischaracterization accounts for 44.1%, and invention accounts for 40.4%.

The labels contain substantial disagreement. Seven responses received four annotations because of assignment collisions in the data collection (Mickus et al., 2026, Sec. 4). The agreement analysis excludes their 99 spans, leaving 12,795 of the 12,894 spans on a uniform three-annotation grid. Of these, 71.7% were marked by one annotator, 20.8% by two, and 7.5% by all three. In addition, 42.2% of hallucinated responses contain no span chosen by at least two annotators. Mean agreement is higher for OCR at 0.578 and miscounting at 0.552 than for mischaracterization at 0.432 or invention at 0.446. We treat this disagreement as an object of analysis instead of discarding it as noise. A check of span boundaries supports this decision. Only 0.6% of marked words have probabilities that vary within the word.

3 System

Candidate extraction. We use spaCy to extract noun phrases, numbers, quoted strings, and adjectives while preserving the original offsets. Local candidates cover 64.5% of gold characters. Adding sentence candidates raises the candidate recall ceiling to 99.8%. Both kinds are sent to the verifier. The final system uses the sentence verdicts as evidence for the decision about the whole response but emits only local candidates, because broad spans flatten the ranking among characters. The 64.5% figure therefore bounds what the emitted spans can cover, not what the verifier sees.

VLM verification. For each candidate, we mark its position inside the full response. We provide the image, original prompt, and marked response to Qwen2.5-VL (Bai et al., 2025). The verifier reports whether the image supports the candidate using `yes` or `no`. It also reports an error category and confidence in its verdict. We compare the 7B and 32B checkpoints without fine-tuning either one. Both use the same prompt, parsing logic, and output budget of 256 tokens. The verifier judges candidates proposed by our code. It does not produce the submission offsets itself.

Confidence signals. All signals are oriented so that larger values indicate greater hallucination suspicion. If the model verbalizes confidence c in its verdict, Signal 1 is

$$s_1 = \begin{cases} c, & \text{verdict=no,} \\ 1 - c, & \text{verdict=yes.} \end{cases} \quad (1)$$

This transformation only aligns the direction of verbalized confidence with the task. It does not assume that s_1 is already calibrated. Signal 2 reads the distribution at the verdict token.

$$s_2 = \frac{P(\text{no})}{P(\text{yes}) + P(\text{no})}. \quad (2)$$

Signal 3, s_3 , is the fraction of `no` verdicts among five generations sampled with temperature. We also average the continuous s_1 and s_2 values over the five draws. The submitted probabilities are raw signal values. Isotonic regression fitted on the calibration split appears only in the calibration analysis of Section 5.1 and is not part of the submitted system.

Aggregation and empty prediction. Candidate scores are projected back to characters. When local candidates overlap, each character receives the largest score. The initial system averages available values of s_1 , s_2 , and s_3 . It emits spans only when at least 60% of candidate verdicts are `no`. This decision has a large effect on the score. The official scorer assigns Cor 1 to an empty prediction on a clean response and Cor 0 to any nonempty prediction on a clean response. A baseline that always returns no span therefore obtains 0.2679 development Cor.

After finding that the system returns too many empty predictions, we retain the cached verifier outputs and replace the single-feature emission rule. The revised method uses a regularized logistic model over summaries from five judgments for each response. We select models and thresholds through cross validation that keeps images grouped on the calibration set. The selected aggregation uses the five-draw average of s_2 alone, with zero weight on s_1 and s_3 , a candidate threshold of 0.5, and only local candidates. Each emitted span carries the error category stated by the verifier. A flagged span with no stated category is labeled *other*. We first retain configurations within 0.005 of the best calibration Cor, then choose the one with the best Cor+Lbl. Test labels are not available and are never used.

4 Experiments

For a candidate, we define its agreement level as the maximum probability of any overlapping gold span, or zero when there is no overlap. We compare candidates at levels 1/3 and 1 using AUROC, where 0.5 represents chance. Verdict accuracy treats any

Model	μc	Verdict acc.	s_1 AUC	s_2 AUC
7B	.924	.538	.518	.595
32B	.973	.744	.674	.712

Table 1: Mean stated confidence (μc), candidate verdict accuracy, and AUROC for ranking candidates overlapping unanimous above singleton gold spans using verbalized (s_1) and logprob (s_2) signals.

gold overlap as positive, while soft ECE uses the mean gold character probability inside the candidate. Error-type accuracy compares the verifier’s stated category with the label of the overlapping gold span that has the highest agreement.

All hyperparameters, including the isotonic mappings used in the calibration analysis, are selected on calibration data and evaluated on development data. We estimate uncertainty with 2,000 paired bootstrap replicates that resample whole images with replacement. All responses and candidates that share an image move together. The 7B and 32B comparisons based on one judgment use identical greedy inference. The experiment with five judgments uses temperature sampling.

5 Results and Analysis

5.1 Scale reveals useful confidence

Table 1 separates the magnitude of stated confidence from its usefulness. Both models are overconfident. The 7B model reports mean confidence of 0.924, while its candidate verdict accuracy is 0.538. The 32B model reports 0.973 confidence and reaches 0.744 accuracy. High raw confidence alone does not indicate a useful uncertainty signal.

The ranking by annotator agreement gives a different result. At 7B, verbalized confidence is close to random when distinguishing candidates overlapping unanimous spans from those overlapping singleton spans, with 0.518 AUROC. Token probability is somewhat more informative at 0.595. At 32B, the same signals reach 0.674 and 0.712. Their mean hallucination scores also rise from candidates with no gold overlap to those overlapping spans selected by one, two, and three annotators. Compared with 7B, the 32B verifier is more confident and ranks candidates by annotator agreement more reliably.

Raw confidence remains poorly calibrated to the soft labels, consistent with the broader calibration literature (Guo et al., 2017). Soft ECE is 0.394 and 0.390 for 7B s_1 and s_2 . The corresponding values

Signal	AUROC	Soft ECE	Values
$\bar{s}_1$	.728	.212	205
$\bar{s}_2$	.737	.197	998
s_3	.716	.230	11

Table 2: Five-draw signals on the common analyzable development cohort of 44,903 candidates. AUROC ranks candidates overlapping unanimous spans above singleton spans. The last column counts distinct signal values.

for 32B are 0.205 and 0.203. Isotonic regression fitted on the calibration set reduces development ECE to values between 0.003 and 0.006. Increasing scale improves discrimination, while explicit recalibration is still needed.

5.2 Averaging improves AUROC, not task score

On development data, the five-draw mean token probability reaches 0.737 AUROC, versus 0.712 for one greedy judgment. Table 2 compares the five-draw signals on a common cohort. Mean token probability has the highest AUROC and lowest soft ECE. Consistency, s_3 , reaches 0.716 AUROC but takes only 11 values and has 0.230 soft ECE.

A stricter paired comparison includes 6,715 candidates from 391 images. Averaging five token probabilities improves AUROC by 0.0163 with a 95% bootstrap interval of [0.0035, 0.0337]. The official development Cor is 0.3061 with one judgment and 0.3059 with the average of five judgments. The paired difference is -0.0003 with an interval of $[-0.0091, 0.0080]$. Cor+Lbl changes by -0.0027 with an interval of $[-0.0076, 0.0025]$. Repeated inference improves ranking among candidates but not the complete decision pipeline. A draw-count ablation shows that three judgments capture 92% of the AUROC gain from one to five sampled draws, with draws four and five adding 0.0009.

5.3 The initial score is dominated by empty outputs

The initial 32B system scores 0.971 Cor on 314 clean development responses but only 0.063 on 858 hallucinated responses. Of the aggregate 0.306 Cor, clean responses contribute 0.260 and hallucinated responses contribute 0.046. Correct empty predictions therefore account for approximately 85% of the score. The rule returns an empty prediction for 644 of 858 hallucinated responses, which leaves hallucination recall at 24.9%. The system

System	Cor	Cor+Lbl	IoU
Always empty	.2679	.2679	.2679
Initial 32B, $k=1$	.3061	.2815	.3343
Five judgments	.3059	.2788	.3384
Revised cached output	.3481	.2875	.3433

Table 3: Development results under the official scorer. The final revision changes aggregation and the decision for each response. Qwen judgments remain unchanged.

Gold label	n	Acc.	Most common prediction
Invention	1,688	.010	Mischaracterization (.899)
Mischaracterization	898	.920	Mischaracterization (.920)
OCR	129	.705	OCR (.705)
Miscounting	322	.947	Miscounting (.947)
Other	219	.151	Mischaracterization (.703)
All	3,256	.391	Mischaracterization (.781)

Table 4: Verifier error-type accuracy for thresholded local candidates that overlap gold spans. The final column gives the most frequent prediction and its within-row rate.

emits 2,893 spans, and 36.1% overlap a gold span. Character precision is 32.7%, and recall is 23.8%.

The response-level decision therefore removes candidate scores before they reach the submission, which is why better candidate AUROC does not automatically improve Cor.

5.4 Revision using cached outputs and final submission

The revised model increases the number of hallucinated development responses receiving any span from 214 to 458 of 858. Correct empty predictions on clean responses decrease from 305 to 281 of 314. With this tradeoff, Cor rises by 0.0419 with an interval of [0.0264, 0.0577]. Cor+Lbl rises by 0.0060 with an interval of [−0.0050, 0.0174], and IoU rises by 0.0091 with an interval of [−0.0042, 0.0223]. The clearest development gain is in Cor.

The smaller Cor+Lbl gain coincides with weak error-type predictions. Table 4 reports local development candidates with $s_2 \geq 0.5$ that overlap a gold span. This conditional analysis does not apply the response-level gate. Overall accuracy is 39.1%, below the 51.8% majority-label baseline. Accuracy is high for miscounting and mischaracterization but only 1.0% for invention. The verifier assigns mischaracterization to 89.9% of invention candidates. Our prompt names the categories without defining their boundaries, which may contribute to this pattern.

Because the analysis conditions on both the score threshold and gold overlap, it diagnoses the

verifier’s category field rather than end-to-end label accuracy. Error typing therefore appears to be one bottleneck for Cor+Lbl, although localization and probability errors also affect that metric.

On the hidden English test set, the system scores 0.3197 Cor, 0.2141 Cor+Lbl, and 0.2905 IoU, ranking 22nd, 22nd, and 18th among 28 systems. The organizer baseline scores 0.1549 on all metrics. The leading Cor system reaches 0.5534 Cor and 0.4336 Cor+Lbl. Against our earlier five-judgment submission, the revision gains 0.0837 Cor and 0.0337 Cor+Lbl. Despite the low rank, the comparison shows that response-level aggregation can discard useful candidate scores.

6 Conclusion

Our experiments suggest that pretrained VLM confidence can support candidate-level hallucination detection. In this setting, discrimination is stronger at 32B than at 7B. Averaging repeated judgments improves candidate ranking, but neither repetition nor low calibration error reliably improves the official score. The gap appears to reflect interactions among several stages. The response-level gate can discard useful candidate scores. Local spaCy candidates restrict the characters eligible for emission. Error-type predictions also appear to limit Cor+Lbl. These observations are specific to one model family, one language, and one benchmark. They motivate evaluating candidate proposal, uncertainty estimation, and aggregation separately.

Limitations

Our findings cover one model family, two scales, one language, and one dataset. They should not be generalized to all VLMs or languages. The comparison controls the model family, but it does not establish that parameter count alone causes the improvement. The experiment with five judgments used two GPU and serving configurations across separate development shards. Bootstrap resampling by image captures sampling uncertainty but not this systematic difference. Analyses of individual candidates inherit errors from heuristic proposals and from the use of character overlap to define verifier correctness. The final aggregation family was selected on calibration data, but the process was exploratory. We inspected an initial candidate optimized for Cor on development data before specifying the final rule that protects

Cor+Lbl. We disclose this order and treat the development result as exploratory rather than as a single confirmatory test. The official rule for clean responses also gives unusual value to empty predictions. Findings about suppression at the response level may differ under another metric.

Candidate proposal is a further limit. Local spaCy candidates bound emitted coverage at 64.5% of gold characters, and sentence candidates recover coverage only at the cost of localization. Three alternatives could relax this bound without fine-tuning the verifier. The first proposes candidates at several granularities, such as clauses and prepositional phrases in addition to noun phrases, and backs off from a rejected sentence to its phrases. The second lets the verifier quote the unsupported words and aligns the quote back to character offsets. The third uses verifier representations or token-level signals to construct finer-grained scores, following the general probing strategy of Kogilathota et al. (2026). Defining the error categories in the prompt is a cheaper change that targets the label problem in Section 5.4.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, et al. 2025. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. [Plausible may not be faithful: Probing object hallucination in vision-language pre-training](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148. Association for Computational Linguistics.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. [Detecting and preventing hallucinations in large vision language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. [Language models \(mostly\) know what they know](#). *arXiv preprint arXiv:2207.05221*.
- Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. 2026. [Halp: Detecting hallucinations in vision-language models without generating a single token](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6067–6085. Association for Computational Linguistics.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics.
- 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*.
- Timothee Mickus, Elaine Zosa, Raúl Vázquez, 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. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [Factscore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. Association for Computational Linguistics.
- Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and

- Hannaneh Hajishirzi. 2024. [Fine-grained hallucination detection and editing for language models](#). In *First Conference on Language Modeling*.
- 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*, pages 10862–10878. 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. Association for Computational Linguistics.
- 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 2472–2497. Association for Computational Linguistics.
- 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*.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). In *International Conference on Learning Representations*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. [Woodpecker: Hallucination correction for multimodal large language models](#). *Science China Information Sciences*, 67(12):220105:1–220105:13.
- For the latter, we average each continuous signal over successfully parsed generations, compute consistency as the fraction of parsed `no` verdicts, and choose the category by majority vote among parsed `no` generations. The verifier output has one line each for verdict, category, and confidence, and parsing uses the final verdict block.
- Candidate extraction uses `spaCy en_core_web_sm` and returns noun phrases, numbers, short quotations, adjectives, and an optional sentence candidate. Offsets are computed directly against the raw response and use half-open character intervals. The data split uses seed 13 and keeps every response associated with an image in the same partition.
- Candidate-level agreement analyses exclude verifier parse failures. The exclusion rates are 0.0% at 7B with one judgment and 5.5% at 32B with one judgment. For the 32B five-judgment run, 92.2% of individual generations parse successfully. These analyses also exclude the seven English samples with four annotations. Official scoring retains every response and uses the organizers’ scorer.
- The selected response gate is an L2-regularized logistic regression with $C = 0.1$. Its target is whether the response contains any gold span. Its standardized input features summarize five-judgment candidate counts, response length, the fraction of `no` verdicts, score distributions, candidate types, and candidate locations. Model-family and threshold selection use image-grouped nested cross validation with five outer folds and four inner folds. Thresholds, feature scaling, and the final logistic model are fitted without test labels. The isotonic mappings reported in Section 5.1 are fitted on the calibration split for analysis only and are not applied to submitted probabilities.

A Reproducibility Details

We use the exact checkpoints `Qwen/Qwen2.5-VL-7B-Instruct` and `Qwen/Qwen2.5-VL-32B-Instruct`, served with vLLM. Each request permits at most 256 generated tokens and asks for top-20 token log probabilities. One-judgment inference is greedy. Five-judgment inference uses temperature 0.7.