

CAML at SHROOM-Visions: Competency-Aware Intermediate-Layer Readouts for Multilingual Hallucination Localization

Jiwon Jeong^{1,*}, Kimin Yun^{2,3,*}, Yuseok Bae²

¹The Catholic University of Korea, Bucheon, South Korea

²Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea

³University of Science and Technology (UST), Daejeon, South Korea

*Equal contribution. †Corresponding author. ‡Work done during a UST internship at ETRI.

Emails: jjo0522@catholic.ac.kr, {kimin.yun, baes}@etri.re.kr

Abstract

We present CAML for multilingual hallucination localization at character level in SHROOM-Visions 2026. The system combines intermediate hidden states and four uncertainty features from a frozen Qwen3.5-4B evaluator, mapping supplied responses to character probabilities and five error categories. Probes from five cross-validation folds use temperature scaling and probability averaging, with about 3.37M trainable parameters in total. We compare readout depths under a common training and evaluation protocol. Selection on validation yields fixed layer 19, with macro character IoU of 0.3704; bootstrap intervals do not resolve its small differences from adjacent depth policies. Separately, the official challenge summary for the team gives macro IoU of 0.3825 and Spearman correlation of 0.4670, averaged over four languages. Independently trained Gemma probes also support localization, with fixed L24 scoring above both tested Gemma submissions combining layers. Across backbones, a single intermediate-layer readout consistently proves effective for localization, although the best depth remains backbone-specific rather than universal.

1 Introduction

Vision-language models can generate fluent descriptions that contain unsupported objects, attributes, text, or counts. For applications such as accessibility and product description, a document-level hallucination label is insufficient, since users need to know where the unsupported content occurs. SHROOM-Visions 2026 (Vázquez et al., 2026) evaluates this problem at character level in four languages on SHEEP (Mickus et al., 2026), which includes both model-generated and human-written responses. Each system predicts a hallucination probability for every character and assigns detected spans to one of five categories, namely

invention, mischaracterization, OCR, miscounting, or other.

Our hypothesis is that a VLM’s intermediate representations already carry signals for locating unsupported content in a supplied response, as suggested by work on OCR heads and probes for OCR errors (Baek et al., 2025; Yao et al., 2025). We ask how to turn these signals into multilingual character-level predictions and how to choose a readout depth.

We present CAML, an ensemble of five temperature-scaled probes on a frozen Qwen3.5-4B evaluator (Qwen Team, 2026), illustrated in Figure 1. During training, the evaluator reads an image with a supplied response, probes are fitted to the hidden states and token uncertainty of candidate layers, and the probes are compared on validation data to select a readout depth. At inference, the ensemble at the selected depth maps the same features to character-level hallucination probabilities and categories. We use *competency-aware* as an operational name for hallucination probing, without claiming explicit self-assessment. Our contributions are (1) a parameter-efficient system mapping intrinsic VLM features to character probabilities and error categories, (2) a controlled comparison of adjacent Qwen depths with matched calibration states and bootstrap uncertainty, and (3) depth comparisons on Qwen2.5-VL-3B (Bai et al., 2025) and Gemma-4-E4B (Gemma Team, 2026), with additional Gemma challenge results.

2 Related Work

SHROOM and Mu-SHROOM established multilingual hallucination span localization (Mickus et al., 2024; Vázquez et al., 2025). Probes of internal states reveal signals of truth and hallucination in language models (Azaria and Mitchell, 2023; Burns et al., 2023; Chen et al., 2024; Dasgupta et al., 2025).

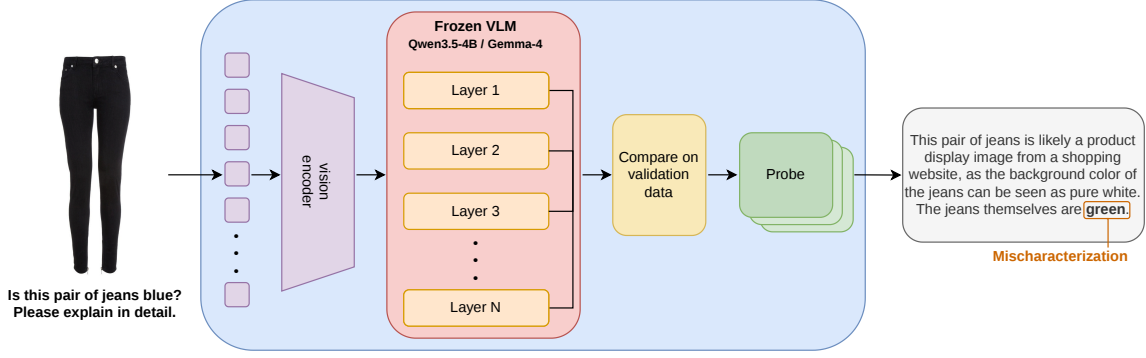


Figure 1: Illustrative walkthrough of our pipeline on a single example. Candidate layers are compared under a fixed rule to select a layer. Its ensemble of five probes flags the hallucinated span (here, “green”) and its category.

Our motivation connects two VLM studies. Baek et al. (2025) characterize OCR attention heads and test their role through interventions. Yao et al. (2025) predict the reliability of OCR answers from hidden states or attention patterns, combining predictions by voting across selected layers. CAML instead averages fold posteriors at a single depth and localizes errors of five categories within a supplied response.

HalLoc (Park et al., 2025) provides hallucination annotations for each token and localization models. HALP (Kogilathota et al., 2026) predicts hallucination risk before VLM generation and finds that informative layers vary by architecture. CAML evaluates a supplied response after teacher forcing, preserving multilingual character offsets and predicting six classes.

3 System

3.1 Task formulation

Given an image, a prompt, and a supplied response R , the system assigns each character of R a hallucination probability and, for detected spans, one of five error categories. The evaluator, however, operates on subword tokens, which do not align one to one with characters. A single character, particularly in Chinese, may be split across several byte-level tokens, and one token may cover several characters. We therefore tokenize R , record each token j ’s character span $[a_j, b_j)$, and group characters by the set of tokens covering them, $C_s = \{j : a_j \leq c < b_j\}$ for a character c . Each maximal run of characters sharing the same C_s forms a *segment*. Characters that no token span covers, such as gaps between adjacent offsets, are attached to a neighboring token, so the segments partition R exactly and every character receives a

score. A segment is thus an alignment unit rather than a token or a linguistic chunk.

3.2 Layer readouts and probe

For the primary Qwen3.5-4B system, features come from a single forward pass over the image, the prompt, and the supplied response concatenated as one sequence, without generating any tokens. For each response token r_j we read its hidden state $h_j^{(\ell)} \in \mathbb{R}^{2560}$ at decoder layer ℓ (one-based) and the next-token distribution P_j that the evaluator predicts at position $j - 1$, before seeing r_j . From P_j we form a four-dimensional uncertainty vector

$$u_j = [\pi_j, -\log \pi_j, \max_v P_j(v), H(P_j)], \quad (1)$$

where $\pi_j = P_j(r_j)$ is the probability assigned to the target token, followed by its negative log-likelihood, the maximum vocabulary probability, and the entropy. Averaging $h_j^{(\ell)}$ and u_j over the tokens in C_s and concatenating the two means gives the segment feature $x_{i,s}^{(\ell)} \in \mathbb{R}^{2564}$, where i indexes examples. We train only a small probe $g_{f,\ell}$ for each cross-validation fold f and layer ℓ , and store its standardization statistics, temperature, and language-specific threshold together as one calibration state.

Each probe $g_{f,\ell}$ is an MLP with hidden widths 256 and 64, GELU activations (Hendrycks and Gimpel, 2016), and dropout (Srivastava et al., 2014). Its six outputs form one mutually exclusive distribution rather than six independent detectors, where class 0 is *non-hallucinated* and classes 1–5 are invention, mischaracterization, OCR, miscounting, and other. The five fold probes together contain about 3.37M trainable parameters, excluding the frozen backbone. Writing $z_{i,s}^{(f,\ell)} =$

$g_{f,\ell}((x_{i,s}^{(\ell)} - \mu_{f,\ell})/\sigma_{f,\ell}) \in \mathbb{R}^6$ for the logits, the temperature-scaled class distribution and the segment hallucination probability are

$$q_{i,s}^{(f,\ell)} = \text{softmax}\left(\frac{z_{i,s}^{(f,\ell)}}{T_{f,\ell,\lambda}}\right), \quad p_{i,s}^{(f,\ell)} = 1 - q_{i,s,0}^{(f,\ell)}, \quad (2)$$

where $\mu_{f,\ell}$ and $\sigma_{f,\ell}$ are standardization statistics from the training fold and $T_{f,\ell,\lambda}$ is a temperature fitted for that fold and language λ . A language-specific threshold $\tau_{f,\ell,\lambda}$, fitted on the same data, determines when $p_{i,s}^{(f,\ell)}$ counts as a detection.

3.3 Ensemble with matched calibration

CAML is an ensemble of the five fold probes. Given depths $\ell = (\ell_0, \dots, \ell_4)$ assigned to the folds, it averages the five temperature-scaled class posteriors and takes the complement of the non-hallucinated class,

$$\bar{q}_{i,s,k}(\ell) = \frac{1}{5} \sum_{f=0}^4 q_{i,s,k}^{(f,\ell_f)}, \quad \bar{p}_{i,s} = 1 - \bar{q}_{i,s,0}, \quad (3)$$

and likewise averages the language thresholds of the same probes,

$$\bar{\tau}_\lambda(\ell) = \frac{1}{5} \sum_{f=0}^4 \tau_{f,\ell_f,\lambda}. \quad (4)$$

Changing a fold’s depth ℓ_f swaps its probe, standardization statistics, temperature, and threshold as one unit, so no hidden state is ever scored with calibration parameters fitted to a different layer.

Finally, the segment-level outputs are mapped back to characters. Every character in segment s receives hallucination probability $\bar{p}_{i,s}$ and category $\arg \max_{k \in \{1, \dots, 5\}} \bar{q}_{i,s,k}$, and is marked positive when $\bar{p}_{i,s} \geq \bar{\tau}_\lambda$. Contiguous positive characters with identical scores and categories are exported as one span, preserving the original Unicode offsets without additional smoothing.

3.4 Readout depth selection

We compare three depth policies under the same training and inference protocol. Fixed L18 and fixed L19 assign one layer to all folds, whereas a fold-wise policy chooses a layer for each fold by maximizing macro IoU on its inner calibration subset (Section 3.5) over the candidate set $\{18, 19, 21\}$. Since only L18 and L19 were chosen, we denote this policy selected L18/L19. On official validation, we adopt fixed L19 as the operational configuration

when it improves macro IoU over selected L18/L19 without reducing any language-level or macro metric by more than 0.005. This rule is a deployment criterion rather than a significance test.

In addition, *Concat* configurations concatenate the hidden states of two layers with the uncertainty vector u_j and feed the result to a single probe in each fold. Their fold probabilities are averaged as in Equation 3, so Concat differs from selected L18/L19 in combining layers within a probe rather than across folds.

3.5 Probe training and calibration

Training data are split into five outer folds with disjoint image groups. Within each outer training partition, one fifth of the data is held out as an inner calibration subset, which serves early stopping, temperature fitting, and threshold selection, while the outer evaluation fold remains untouched. The separate official validation split is used only to compare deployment configurations.

Each probe is trained with cross-entropy weighted by the inverse square root of class frequencies, using AdamW (Loshchilov and Hutter, 2019) with learning rate 10^{-3} and batches of 2,048 segments, for up to eight epochs with early stopping. A segment takes the category whose annotation-probability-weighted character overlap is largest, and segments with no overlap are labeled non-hallucinated.

For each language, we fit the temperature by minimizing cross-entropy with L-BFGS (Liu and Nocedal, 1989) on the calibration subset and select the threshold that maximizes mean per-example character IoU on the same subset.

4 Experiments

4.1 Setup

SHEEP (Mickus et al., 2026) provides 12,076 labeled examples (1,651 image groups) for readout training, partitioned by image group into nested folds with zero group overlap, plus the official validation split (3,026 examples) and challenge split (4,898 examples), each roughly balanced across the four languages. As an external reference we use the participant-kit baseline (Nath et al., 2025), which pairs LLaVA-1.5-7B (Liu et al., 2024) with a shared multilingual six-class MLP, after correcting its span alignment to use tokenizer offsets instead of decoded strings, and evaluate it with the official scorer. Its different backbone and fit-

Table 1: Official validation results. I/S denote character IoU and Spearman correlation (Cor); Macro is the unweighted language mean. Bootstrap 95% intervals (10,000 replicates) for macro differences: fixed L19 minus selected L18/L19, IoU $+0.0020$ $[-.0026, .0066]$ and Cor -0.0007 $[-.0054, .0042]$; selected L18/L19 minus fixed L18, IoU $[-.0057, .0076]$. All include zero.

System	Backbone	EN		FR		IT		ZH		Macro	
		I	S	I	S	I	S	I	S	I	S
Kit baseline (corrected)	LLaVA-1.5-7B	.2778	.2992	.2498	.2676	.2667	.2885	.3591	.3873	.2884	.3106
Fixed L18 (adjacent depth)	Qwen3.5-4B	.3365	.4068	.3493	.4142	.3502	.4155	.4339	.4848	.3675	.4304
Selected L18/L19	Qwen3.5-4B	.3347	.4040	.3404	.4107	.3565	.4255	.4421	.4959	.3684	.4340
Fixed L19 (operational)	Qwen3.5-4B	.3342	.4040	.3502	.4115	.3517	.4207	.4454	.4971	.3704	.4333

Backbone / depth	IoU	Cor
Qwen3.5-4B		
Fixed L18	.3545	.4090
Fixed L19	.3559	.4110
Fixed L21	.3431	.3912
Qwen2.5-VL-3B		
Fixed L23	.3310	.3779
Fixed L25	.3365	.3843
Fixed L27	.3348	.3821
Gemma-4-E4B		
Fixed L21	.3387	.3898
Concat L21/L32	.3448	.3896
Fixed L25	.3485	.3941
Fixed L24	.3509	.3963

Table 2: Out-of-fold candidate comparisons on training data ($n = 12,076$); bold marks the best per backbone.

ting recipe make this a whole-system rather than a probe comparison. We report character IoU between predicted and gold hallucination spans and Cor, the mean Spearman correlation per example against gold character probabilities from multiple annotators. Cor+Lb1 averages this correlation over the categories present in the gold or predicted spans. Human-written SHEEP examples occur only in the challenge split, while training and validation use model-generated responses. We use the post-trained Qwen3.5-4B checkpoint and the instruction-tuned Qwen2.5-VL-3B-Instruct and Gemma-4-E4B-it checkpoints, with 32, 36, and 42 decoder layers.

Research with AI assistance. DeepScientist (Weng et al., 2026) supported exploratory depth sweeps. The authors defined the research question and protocols, selected the reported comparisons, and verified all evidence.

4.2 Validation results

Table 1 shows the complete validation comparison. All three Qwen policies exceed the kit baseline on every reported metric, both per language and

macro, although the controlled comparison is only among the Qwen policies. Fixed L19 has the highest macro IoU (0.3704), while selected L18/L19 has the highest macro Cor (0.4340). The IoU gain of 0.0020 for fixed L19 over selected L18/L19 meets the selection criterion, and its largest decrease, approximately 0.0048 in both Italian IoU and Cor, stays within the 0.005 tolerance. The bootstrap intervals in Table 1 all include zero, so the rule selects L19 as an operational choice among comparably performing adjacent depths.

4.3 Internal depth selection

Table 2 reports out-of-fold scores on the 12,076 training examples, each scored by the probe trained on the other outer folds with inner calibration. For Qwen3.5-4B, a coarse search over 40–80% of decoder depth in 10-point steps found L19 winning most often on inner calibration IoU across seeds and folds, and a predefined refinement at ± 5 points gave the candidate set $\{18, 19, 21\}$ from which selected L18/L19 was drawn. For Qwen2.5-VL, L25 leads the primary run over L23/L25/L27, although the preferred layer varies across seeds and folds. For Gemma, L24 scores above L21, L25, and concat L21/L32, and its three-seed mean IoU (.3476) also exceeds L25 (.3457), motivating its submission.

4.4 Official challenge results

Table 3 reports team-level bootstrap means from the overview (Vázquez et al., 2026) in panel (a) and configuration-level scores in panel (b). Team estimates cannot be attributed to one configuration, and their higher IoU than validation reflects the different aggregation and the inclusion of human-written samples rather than a measured gain. Besides the validation-selected configurations, we submitted two diagnostic-motivated variants, Qwen concat L16/L19 from onset layers and Gemma concat L33/L34 from sampled restoration peaks for

(a) Team-level overview				(b) Individual submissions		
Language	Cor+Lbl	Cor	IoU	Submission	Cor+Lbl	Cor
EN	.338	.425	.337	Qwen fixed L19	.37085	.45638
FR	.375	.482	.399	Qwen selected L18/L19	.37625	.46000
IT	.379	.483	.386	Qwen concat L16/L19	.37485	.46015
ZH	.422	.478	.408	Gemma fixed L24	.33693	.41998
Macro	.3785	.4670	.3825	Gemma concat L21/L32	.33275	.41283
				Gemma concat L33/L34	.29173	.35288

Table 3: Official challenge results. (a) Team caml’s score means from the overview’s 25,000 bootstrap resamples (Vázquez et al., 2026, Tables 4–7); macro averages its rounded language scores. (b) Submission scores averaged over four languages. IoU for individual configurations and paired significance tests are unavailable.

OCR/color and counting. Full rankings and cross-team analysis are given in the overview.

Two observations follow from panel (b). Among Qwen submissions, selected L18/L19 exceeds fixed L19 on Cor+Lbl and Cor while concat L16/L19 attains the highest Cor, so the configuration selected by validation IoU is not the one with the highest challenge correlation. For Gemma, whose probes are trained independently of Qwen, fixed L24 exceeds both submitted concatenations on Cor+Lbl and Cor.

4.5 Diagnostic: where count signals emerge

To ask why an intermediate depth works, we located where count-sensitive information first appears in Qwen3.5-4B by activation patching. On 160 synthetic pairs of colored circles differing by one circle of the queried color, we restored clean-image activations at layers 13–25 against a pixel-matched control edit, and define *onset* as the earliest layer with a consistently positive restoration effect for both adding and removing a circle. Counting onset falls at L19, the depth validation independently selected, while a parallel color-substitution experiment places onset at L16, the layer paired with L19 in the concat submission. Whether such onsets predict localization performance remains open and motivates controlled hallucination experiments across error types.

5 Conclusion

CAML localizes multilingual hallucinations at character level with frozen VLM features and lightweight probes. Across Qwen3.5-4B, Qwen2.5-VL-3B, and Gemma-4-E4B, a single intermediate-layer readout consistently supports localization and combining two layers brought no consistent gain. The best depth is backbone-specific, with validation selecting Qwen L19 and Gemma probes favor-

ing L24, while adjacent depths perform comparably, so readout depth should be selected and calibrated within each backbone rather than transferred. The onset diagnostic offers a first explanation, since count-sensitive signals emerge at L19, the layer validation chose, and controlled hallucination experiments across error types will test whether such onsets predict localization performance.

Limitations

Our experiments identify effective intermediate readouts on three VLM backbones but do not explain which internal signals the probes exploit. The onset diagnostic covers counting and color only, and whether onsets predict localization performance remains untested. A systematic study with controlled hallucinations across categories and backbones is left to future work.

Ethical Considerations

AI assistance supported exploratory experiments only. The authors selected and verified all reported evidence and added no AI-generated annotations to the benchmark.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities) and the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (26CB1200, Development and Application of Science-Specialized Multimodal Foundation Models).

References

- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. 2025. [How do large vision-language models see text in image? unveiling the distinctive role of OCR heads](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20441–20453, Suzhou, China. Association for Computational Linguistics.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-VL Technical Report](#). *Preprint*, arXiv:2502.13923.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *The Eleventh International Conference on Learning Representations*.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. [INSIDE: LLMs’ Internal States Retain the Power of Hallucination Detection](#). In *The Twelfth International Conference on Learning Representations*.
- Sharanya Dasgupta, Sujoy Nath, Arkaprabha Basu, Pourya Shamsolmoali, and Swagatam Das. 2025. [HalluShift: Measuring distribution shifts towards hallucination detection in LLMs](#). In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Gemma Team. 2026. [Gemma 4 Technical Report](#). *Preprint*, arXiv:2607.02770.
- Dan Hendrycks and Kevin Gimpel. 2016. [Gaussian error linear units \(GELUs\)](#). *Preprint*, arXiv:1606.08415.
- 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, Rabat, Morocco. Association for Computational Linguistics.
- Dong C. Liu and Jorge Nocedal. 1989. [On the limited memory BFGS method for large scale optimization](#). *Mathematical Programming*, 45:503–528.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *The Seventh International Conference on Learning Representations*.
- 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](#). *Preprint*, arXiv:2608.01021. Version 1.
- 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.
- Sujoy Nath, Arkaprabha Basu, Sharanya Dasgupta, and Swagatam Das. 2025. [HalluShift++: Bridging language and vision through internal representation shifts for hierarchical hallucinations in MLLMs](#). In *Proceedings of the Sixteen Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9. Association for Computing Machinery.
- Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. [HallLoc: 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.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#). Qwen Blog. <https://qwen.ai/blog?id=qwen3.5>.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: A simple way to prevent neural networks from overfitting](#). *Journal of Machine Learning Research*, 15(56):1929–1958.
- 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: 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, Vienna, Austria. 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 Vaiani, Jörg Tiedemann, and Timothee Mickus. 2026.

Overview of SHROOM-Visions 2026: A Shared Task on Hallucination Detection in Large Vision-Language Models. *Preprint*, arXiv:2608.25662. Version 2.

Yixuan Weng, Minjun Zhu, Qiujie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. 2026. [Deepscientist: Advancing frontier-pushing scientific findings progressively](#). In *The Fourteenth International Conference on Learning Representations*.

Jihan Yao, Achin Kulshrestha, Nathalie Rauschmayr, Reed Roberts, Banghua Zhu, Yulia Tsvetkov, and Federico Tombari. 2025. [Reading Between the Lines: Abstaining from VLM-Generated OCR Errors via Latent Representation Probes](#). *Preprint*, arXiv:2511.19806. Version 1.