

SmurfCat at SHROOM-Visions: Synthetic Augmentation for Multilingual Span-Level Hallucination Detection

Weerathep Rattanajaratkul¹ Alexander Panchenko^{1,2} Elisei Rykov¹

¹Skolkovo Institute of Science and Technology, ²AIRI

{weerathep.rattanajaratkul, a.panchenko, elisei.rykov}@skoltech.ru

Abstract

We present our system for the SHROOM-Visions shared task: character-level hallucination localization and five-category classification in image-grounded responses across four languages. We investigate a synthetic-data augmentation pipeline combining annotator-agreement- and class-rarity-aware upsampling with controlled, image-grounded hallucination injection, used to fine-tune Qwen3-VL-8B-Instruct with LoRA. A conservative augmentation configuration achieves the best validation Cor+Lbl and significantly outperforms our initial, more aggressive configuration, though its gain over the original data alone is not significant and transfers to the hidden test set only for Chinese. Per-category analysis shows it helps invention and miscounting but consistently degrades OCR, indicating uneven transfer to the categories it targets. Our submissions placed 7th of 28 (English), 5th of 21 (French), 6th of 21 (Italian), and 5th of 22 (Chinese) under Cor+Lbl. The code is available on GitHub.¹

1 Introduction

Vision-language models frequently produce text that is fluent but not grounded in the accompanying image: they invent objects, misdescribe visible content, misread on-image text, or miscount visible items. The SHROOM-Visions shared task (Vázquez et al., 2026) asks participants to localize such hallucinations at the character level and assign each span one of five categories, over LVLM-generated and human-written responses to image-grounded questions in four languages (Mickus et al., 2026). This is more demanding than simply deciding whether a response contains a hallucination: a system must identify the exact unsupported substring and determine the nature of the error.

The task releases labeled training data and a hidden test set whose gold annotations are held by

the organizers. We carve out an internal validation split of 3,020 examples from the training data and use the remaining 12,082 for training; final results come from the organizers’ evaluation platform. The available supervision is strongly imbalanced across categories, and synthetic augmentation is a natural remedy, but the multimodal setting makes it easy to generate corruptions that are artificial or insufficiently grounded in the image.

We design a synthetic-data pipeline combining annotator-agreement and class-rarity-aware upsampling with controlled hallucination injection grounded in image and response signals, and tune the strength of rare-category upsampling to reduce spurious predictions. We use it to fine-tune Qwen3-VL-8B-Instruct with LoRA (Hu et al., 2022) as an end-to-end detector that takes an (image, question, response) triple and directly generates localized, categorized hallucination spans. Our contributions:

- A grounded synthetic-data augmentation pipeline for multilingual multimodal hallucination detection, combining agreement- and rarity-aware upsampling with controlled error injection based on real image and response signals.
- A competitive model trained with this pipeline to directly produce localized and categorized hallucination spans across four languages.

2 Related Work

Hallucination detection has been the focus of several recent shared tasks, though prior efforts were text-only. Mu-SHROOM formulated it as multilingual span labeling over LLM outputs across 14 languages (Vázquez et al., 2025); SHROOM-Visions extends this to the multimodal setting, where spans must be verified against visual evidence and assigned error categories.

¹<https://github.com/s-nlp/shroom-visions>

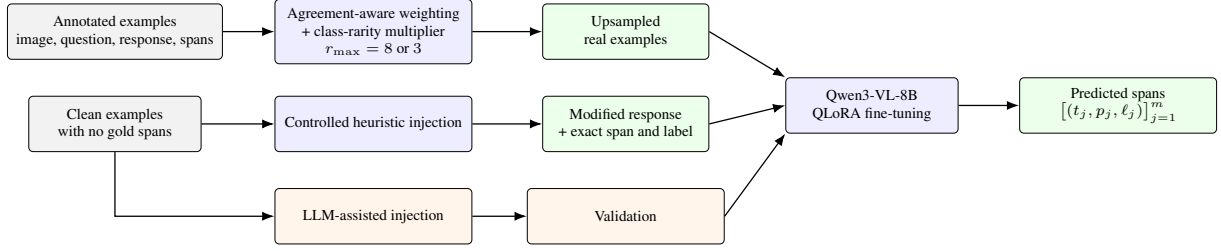


Figure 1: Overview of the data-augmentation and fine-tuning pipeline. Here, t_j is an exact substring of the evaluated response, p_j is its predicted hallucination probability, and ℓ_j is its category label.

The scarcity of labeled spans led several strong Mu-SHROOM submissions to synthetic supervision: Rykov et al. (2025) combined a fine-tuned span detector with uncertainty-based predictions (top three in four tracks); Kobus et al. (2025) and Mo et al. (2025) also fine-tuned on synthetic data, with particularly strong Spanish and top-five results, respectively; others relied on training-free strategies such as translation-based prompting (Karkani et al., 2025). We revisit synthetic augmentation in the multimodal setting, where generated errors must remain grounded in the image and controlled across categories.

3 Data Augmentation

The training data are strongly imbalanced across categories: OCR, other, and miscounting occur in 6.0%, 8.9%, and 12.3% of examples, versus 41–44% for invention and mischaracterization. Our augmented configuration combines the original data with two components that increase minority-category representation without discarding annotation reliability: agreement- and rarity-aware up-sampling, and grounded heuristic injection. Figure 1 summarizes the pipeline, including the LLM-assisted extension (Section 3.3).

The design of these components was informed by a preliminary screening of augmentation strategies using a smaller backbone, reported in Appendix A.

3.1 Agreement- and Rarity-Aware Upsampling

We duplicate real examples by both annotation certainty and class frequency: the agreement component assigns larger duplication factors to spans marked consistently by more annotators, and a per-category rarity multiplier is derived from the ratio between the most common category’s frequency and each minority category’s.

The maximum rarity multiplier is a hyperparameter. A cap of $8\times$ leaves every category’s ratio-derived multiplier uncapped, so we treat it as our initial configuration; we additionally evaluate a more conservative $3\times$ cap, chosen after diagnosing a label over-triggering effect in the $8\times$ configuration (Section 5.2), which limits how much rare labels can dominate the augmented distribution.

3.2 Grounded Heuristic Injection

We construct additional examples by introducing controlled errors into clean responses, each grounded in the paired image or the response’s own content rather than unconditional text corruption.

For OCR, we apply visually plausible character substitutions to text an OCR model detects in the image; for miscounting, we shift a number mentioned in the response, preferring spelled-out forms; mischaracterization and other substitute language-specific attributes (colour, size, material, state, direction); invention inserts a short object mention at a random sentence boundary. Appendix E compares injected and natural span lengths per category; they are close for OCR and miscounting but diverge for the other three, which we return to in Section 5.2.

3.3 LLM-Assisted Extension

We additionally evaluate an LLM-assisted extension targeting OCR, other, and miscounting: a commercial LLM generates a modified response containing one hallucination of the requested category, a separate validation call checks whether the injected claim is factually inconsistent with the image, and candidates are retained only if confirmed.

This component is evaluated as an optional extension to the grounded augmentation configuration rather than included by default.

4 Experimental Setup

4.1 Training Configurations

We compare a baseline trained on the 12,082 original examples with two systems using our augmentation pipeline, whose design was informed by preliminary experiments in Appendix A.

The original augmented configuration uses $r_{\max} = 8$ and all three methods (upsampling, heuristic injection, LLM-assisted injection), yielding 34,243 training examples: 12,082 original, 18,423 upsampled, 2,538 heuristic, and 1,200 LLM-generated.

The tuned configuration reduces $r_{\max}$ to 3, retains heuristic examples, and excludes LLM-assisted injection (removed because it reduced validation Cor+Lbl), so the two submitted systems differ in both respects, confounding their direct comparison. We therefore additionally evaluate a third configuration, *tuned + LLM* ($r_{\max} = 3$ with LLM-assisted injection): comparing it against the original configuration isolates the rarity-cap change (holding LLM-assisted injection on), and comparing it against tuned isolates the LLM-assisted-injection change (holding $r_{\max} = 3$); we did not train the remaining cell ($r_{\max} = 8$, no LLM), so the decomposition in Section 5.2 stays partial.

4.2 Model and Fine-Tuning

We fine-tune Qwen3-VL-8B-Instruct (Qwen Team, 2025) using LoRA (Hu et al., 2022) with 4-bit NF4 quantization (QLoRA recipe; Dettmers et al., 2023), with loss computed only over the assistant’s span-JSON tokens. All variants share the same model and training configuration, differing only in training data (hyperparameters in Appendix B).

4.3 Input and Output Formulation

Each example consists of an image, a question, and the response to evaluate, preceded by a fixed instruction defining the task and five categories. The target is a JSON array of {text, prob, label} objects, with text extracted via gold character offsets so the model learns to reproduce exact hallucinated substrings. This treats hallucination detection as direct structured generation rather than token classification, jointly identifying text, confidence, and category.

4.4 Model Selection

For each language, we select the submission model based on performance on our internal validation

split under the official metrics, before evaluation against the hidden test annotations; selected models are reported with the test results.

5 Results

5.1 Validation Performance

Configuration	$r_{\max}$	EN	FR	IT	ZH
Baseline (original data)	–	0.362	0.395	0.413	0.495
Original augmentation	8	0.364	0.380	0.415	0.483
Tuned augmentation	3	0.373	0.394	0.418	0.498
Tuned + LLM-assisted	3	0.383	0.399	0.396	0.480

Table 1: Validation Cor+Lbl by language and training configuration. Bold indicates the best score in each language.

Original augmentation improves IoU (0.438→0.453) and Cor (0.471→0.488) but reduces Cor+Lbl (0.417→0.411): better label-agnostic localization, less accurate category-specific predictions. Tuned augmentation achieves the best IoU (0.458) and Cor+Lbl (0.422). A paired bootstrap (10,000 resamples over the 3,020 validation examples) shows tuned’s gain over original augmentation is significant ($\Delta=+0.010$, 95% CI [0.001, 0.019]), but its gain over the baseline is not ($\Delta=+0.004$, CI [−0.004, 0.013]; see Limitations for what this bootstrap does and does not capture).

5.2 Mechanism and Per-Category Analysis

The $r_{\max} = 8$ configuration raises Cor but lowers Cor+Lbl because it predicts $2.5\times$ more *spurious* labels per example than the baseline, consistent with the multiplier training the model to over-expect rare classes it does not observe (full statistics in Appendix C). This motivated the lower $r_{\max} = 3$ cap.

Category	Base.	Orig.	Tuned	+LLM
OCR	0.358	0.274	0.308	0.293
other	0.027	0.092	0.068	0.114
miscounting	0.392	0.398	0.422	0.405
invention	0.204	0.205	0.221	0.214
mischaracterization	0.118	0.146	0.138	0.134

Table 2: Per-category validation Cor (the label-specific term underlying Cor+Lbl), aggregated over languages, for the baseline, original augmentation ($r_{\max}=8$, +LLM), tuned augmentation ($r_{\max}=3$, no LLM), and tuned + LLM-assisted ($r_{\max}=3$, +LLM). Bold marks the best configuration per category.

Table 2 tests our motivating hypothesis directly: miscounting, invention, and *other* improve under

some augmented configuration, but OCR is *worse* under all three than the baseline – the opposite of its intended effect. The tuned + LLM-assisted row of Table 1 further isolates the tuned configuration’s two changes, which pull in *opposite* directions across languages rather than acting uniformly (deltas in Appendix C); we did not train the remaining cell ($r_{\max} = 8$, no LLM), so the decomposition stays partial. Appendix D shows the tuned configuration’s Chinese gain traces to invention and miscounting, not OCR.

Configuration	$r_{\max}$	EN	FR	IT	ZH
Leaderboard best	–	0.434	0.474	0.451	0.504
Baseline (original data)	–	0.352	0.397	0.393	0.450
Original augmentation	8	0.336	0.376	0.369	0.418
Tuned augmentation	3	0.337	0.378	0.389	0.456
Tuned + LLM-assisted	3	0.335	0.349	0.386	0.438

Table 3: Test-set Cor+Lbl by language and training configuration, with the best official leaderboard score shown for reference. Bold indicates the best result among our systems in each language.

5.3 Test Performance

Table 3 reports test-set Cor+Lbl, with the best leaderboard score for reference. Tuned augmentation outperforms original augmentation and tuned + LLM-assisted in all four languages (confirming our exclusion of LLM-assisted injection), mirroring the significant validation-set gain above, most for Italian (+0.020) and Chinese (+0.038).

Compared with the baseline, tuned augmentation is higher for Chinese (0.456 vs. 0.450), slightly lower for Italian (0.389 vs. 0.393), and behind for English and French; the gap to the leaderboard best is smallest for Chinese (0.048) and Italian (0.058). Since the platform reports only aggregate metrics, our per-category analysis remains validation-only, where the Chinese gain traces to invention and miscounting rather than OCR.

Our official submission ranked 7th of 28 teams for English, 5th of 21 for French, 6th of 21 for Italian, and 5th of 22 for Chinese under Cor+Lbl.

6 Conclusion

We presented a grounded synthetic-data augmentation pipeline for multilingual multimodal hallucination detection, used to train a competitive Qwen3-VL-8B-Instruct model across four languages. Its effect is modest and uneven: the tuned configuration’s validation gain over the baseline is not significant, though its gain over our initial, more

aggressive configuration is. It helps invention and miscounting but consistently *degrades* OCR, and its two component changes pull in opposite directions across languages rather than acting uniformly. On the test set it improved over both the original and tuned + LLM-assisted configurations in every language but exceeded the baseline only for Chinese. Our official submission ranked between 5th and 7th under Cor+Lbl.

Future work should complete the factorial ablation design, compare the lexical form and image-question context of synthetic and natural hallucinations beyond span length, and test whether a classifier can distinguish the two.

Limitations

Our preliminary augmentation screening used a smaller backbone (Qwen2-VL-2B) and internal metrics, so its negative results do not establish that discarded strategies would underperform with Qwen3-VL-8B under the official protocol.

Our ablation (Section 5.2) isolates the rarity-cap and LLM-assisted components using three of the four cells of a 2×2 design; we did not train the fourth ($r_{\max} = 8$, no LLM), so the decomposition is incomplete. Since each configuration was fine-tuned only once, our bootstrap intervals capture example-level variance only, not training-seed variance.

Our per-category analysis and span-length comparison (Appendix E) use the validation split, since the organizers do not release test-set gold labels; we cannot verify the same pattern holds on hidden data, and we did not compare lexical form, image-question context, or test whether a classifier can distinguish synthetic from natural hallucinations.

Finally, the tuned configuration exceeds the baseline only for Chinese; this reflects this particular combination of methods on a single training run per language, not evidence that synthetic augmentation is inherently language-dependent for this task.

References

- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Dimitra Karkani, Maria Lymperaiou, George Filandrianos, Nikolaos Spanos, Athanasios Voulodimos, and Giorgos Stamou. 2025. [AILS-NTUA at SemEval-2025 task 3: Leveraging large language models and translation strategies for multilingual hallucination detection](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1289–1305, Vienna, Austria. Association for Computational Linguistics.
- Catherine Kobus, Francois Lancelot, Marion-Cecile Martin, and Nawal Ould Amer. 2025. [ATLANTIS at SemEval-2025 task 3 : Detecting hallucinated text spans in question answering](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1098–1107, Vienna, Austria. 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](#). *Preprint*, arXiv:2608.01021.
- Xinyuan Mo, Nikolay Vorontsov, and Tiankai Zang. 2025. [Team cantharellus at SemEval-2025 task 3: Hallucination span detection with fine tuning on weakly supervised synthetic data](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1724–1736, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3-vl technical report](#). *CoRR*, abs/2511.21631.
- Elisei Rykov, Valerii Olisov, Maksim Savkin, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. [SmurfCat at SemEval-2025 task 3: Bridging external knowledge and model uncertainty for enhanced hallucination detection](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1034–1045, Vienna, Austria. Association for Computational Linguistics.
- Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgrén, 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-vision 2026: A shared task on hallucination detection in large vision-language models](#). *Preprint*, arXiv:2608.25662.

A Preliminary Augmentation Screening

Before developing our final method, we screened six augmentation and resampling strategies using an earlier data release, QLoRA-tuned Qwen2-VL-2B-Instruct, and locally implemented Span IoU, Spearman correlation, and macro category-F1 metrics, each evaluated against the same real-data-only baseline. These scores are not directly comparable to the main experiments, since the backbone, data release, and evaluation implementation differ; the pilot was meant only to identify promising directions.

A.1 Candidate Strategies

Heuristic injection. Clean responses are corrupted via unconditional dictionary substitutions (confusable characters, attribute changes, numerical edits), without verifying the error against the paired image.

LLM-assisted injection. A commercial LLM generates one hallucination per clean response for one of the three rarest categories, validated by a second call.

Cross-lingual transfer. English hallucinated spans are marked before translation, then recovered from the marker positions to produce French, Italian, and Chinese examples.

Back-translation. Hallucinated responses are round-tripped through a pivot language to vary surface form while preserving the hallucinated content.

Agreement-aware upsampling. Real examples are duplicated by annotator agreement, with larger factors for more consistently annotated spans.

Multi-model fusion. Hallucinated spans from multiple LVLMs answering the same image-question pair are merged into one denser synthetic response.

Strategy	IoU	ρ	Macro F1
Real data only	0.310	0.350	0.286
Heuristic injection	0.263	0.302	0.230
LLM-assisted injection	0.276	0.318	0.248
Cross-lingual transfer	0.281	0.305	0.244
Back-translation	0.275	0.306	0.256
Agreement upsampling	0.328	0.350	0.310
Multi-LVLM fusion	0.285	0.317	0.249

Table 4: Preliminary comparison of augmentation and resampling strategies using Qwen2-VL-2B-Instruct and an earlier data release. The results are comparable only within this pilot study.

A.2 Pilot Results

Agreement-aware upsampling was the only strategy to improve over the baseline (higher IoU and macro F1, equal Spearman correlation); all content-generation and transfer-based strategies underperformed on every metric.

These results informed, but do not validate, our final design: we retained reliability-aware resampling and redesigned heuristic/LLM-assisted injection around grounded signals, dropping cross-lingual transfer, back-translation, and multi-model fusion. Given the smaller backbone and earlier data release, this ranking may not hold for the final model.

B Implementation Details

Parameter	Value
Base model	Qwen3-VL-8B-Instruct
LoRA rank / α	16 / 16
Quantization	4-bit NF4
Epochs	2
Micro-batch size	1
Gradient accumulation	8
Effective batch size	8
Maximum sequence length	4,096 tokens
Maximum image area	1024 × 28 × 28 pixels

Table 5: Training hyperparameters shared by all final experimental variants.

Training uses the Hugging Face transformers and peft libraries, with LoRA adapters on all attention and MLP projection layers, batch size one with eight-step gradient accumulation, on a single 40GB GPU.

Examples exceeding 4,096 combined image-and-text tokens are skipped rather than truncated, since truncation could remove part of the response while leaving its target span unchanged. Images are capped at 1024 × 28 × 28 pixels, preserving as-

pect ratio.

Data ingestion and identifiers. Each augmented file concatenates original records with two tagged auxiliary streams: upsampled duplicates keep the original identifier with an appended `__upk_hash` suffix, while synthetic examples receive a fresh `synth_type_` identifier naming the injection rule (e.g. `ocrgrnd`, `cntword/cntdigit`, `attrswap/stateswap`, `objadd`). This exact-by-construction tagging enables the comparison in Appendix E. Images are cached in memory per unique filename to avoid redundant decoding across upsampled copies.

The complete instruction used during training and inference is provided in Appendix F.

C Label Over-Triggering and Ablation Detail

This appendix expands the mechanism analysis summarized in Section 5.2.

Label over-triggering. The $r_{\max} = 8$ and baseline models have similar exact label-set match rates (43.8% vs. 42.6%), and the former has fewer missing labels per example (0.584 vs. 0.728, consistent with its Cor gain), but $2.5\times$ more spurious labels (0.195 vs. 0.078) – categories guessed that gold does not contain. Cor+Lbl averages a correlation term per label in the union of gold and predicted labels, so each spurious label adds a near-zero term that label-agnostic Cor does not penalize the same way, explaining why $r_{\max} = 8$ raises Cor while lowering Cor+Lbl.

Partial ablation deltas. Table 1 reports Cor+Lbl for four configurations. Comparing original augmentation to tuned + LLM-assisted isolates the rarity-cap change (8→3, LLM on): +0.019 EN/FR, −0.019 IT, −0.003 ZH. Comparing tuned + LLM-assisted to tuned isolates the LLM-assisted-injection change ($r_{\max} = 3$ fixed): −0.010 EN, −0.005 FR, +0.022 IT, +0.018 ZH. The tuned configuration’s net gain over the original thus sums two effects that cancel for IT/ZH and reinforce for EN/FR, not a single uniformly beneficial change; the fourth cell ($r_{\max} = 8$, no LLM) was not trained, so we cannot confirm this rarity-cap effect holds with LLM-assisted injection off.

D Per-Language and Per-Category Validation Breakdown

Table 6 expands Table 2 by language, for the baseline and tuned configuration. Values are per-record Cor restricted to spans of the given label, matching the per-label term Cor+Lbl averages over.

Lang.	Category	Baseline	Tuned
EN	invention	0.185	0.186
	mischaracterization	0.120	0.144
	OCR	0.426	0.346
	miscounting	0.361	0.423
	other	0.033	0.134
FR	invention	0.182	0.198
	mischaracterization	0.111	0.136
	OCR	0.355	0.308
	miscounting	0.496	0.449
	other	0.035	0.019
IT	invention	0.230	0.244
	mischaracterization	0.138	0.151
	OCR	0.302	0.345
	miscounting	0.414	0.417
	other	0.021	0.103
ZH	invention	0.229	0.274
	mischaracterization	0.101	0.117
	OCR	0.314	0.215
	miscounting	0.326	0.402
	other	0.022	0.012

Table 6: Per-label Cor by language, real-data baseline vs. tuned augmentation, on the validation split.

For Chinese, gains concentrate in invention (+0.045) and miscounting (+0.076), with a small gain for mischaracterization (+0.016); OCR degrades (-0.099) and other is slightly lower (-0.010) – the drivers are the two most frequent categories, not the rare ones the pipeline targets. OCR degrades for every language, and other improves for English/Italian but not French/Chinese, so neither targeted rare category transfers consistently across languages.

E Synthetic vs. Natural Span Length

To check whether heuristic injection (Section 3) produces spans resembling natural hallucinations, Table 7 compares injected and natural (non-upsampled) gold span lengths per category, using the identifier scheme from Appendix B.

OCR and miscounting – grounded most directly in detected image or response content – are closely length-matched to natural spans. The other three diverge: synthetic mischaracterization spans are single-word swaps (median 1) versus natural’s longer descriptive phrases (median 7); synthetic

Category	Natural		Synthetic	
	Mean	Median	Mean	Median
OCR	8.0	5.0	7.7	6.0
miscounting	6.6	3.0	5.8	3.0
invention	27.0	14.0	38.6	51.0
mischaracterization	14.4	7.0	2.9	1.0
other	65.4	9.0	261.8	145.0

Table 7: Character span length (mean, median) of naturally occurring vs. heuristically injected hallucination spans in the training data, by category.

other spans are far longer (median 145 vs. 9), since state/direction substitutions can affect a larger clause; and synthetic invention insertions (median 51) are longer and more uniform than natural ones (median 14). This mismatch is consistent with, but does not fully explain, the uneven transfer in Table 2, since OCR is length-matched yet still degrades. We did not compare lexical form or image-question context (see Limitations).

F Inference Prompt Template

The following instruction is prefixed to every example, followed by Question asked: <prompt> and VLM response to analyse: <response>; the image is a separate multimodal input, and the target is the JSON array itself.

Identify all hallucinated text spans in the vision model’s response.
A hallucination is text that is factually incorrect compared to what is actually visible in the image.

Categories:
- invention: entities or facts completely absent from the image
- mischaracterization: visible content described incorrectly
- OCR: text visible in the image is misread or mistranscribed
- miscounting: wrong quantity of visible items
- other: hallucination not fitting the above

List spans in the order they appear in the response, left to right.

Return ONLY a JSON array, no markdown fences, no extra text:
[{"text": "<exact text copied from the response>",
"prob": <0.0-1.0>, "label": "<category>"}]
If there are no hallucinations, return: []

G LLM-Assisted Injection Prompts

The optional LLM-assisted extension described in Section 3 uses a two-step generate-then-validate procedure for OCR, other, and miscounting hallucinations.

Generation

You are building hallucination-detection training data.

Task: Rewrite the VLM response below to inject EXACTLY ONE hallucination of type '<label>'.
Definition: <category-specific definition>

Rules:

1. Make ONLY ONE factual change; keep everything else identical
2. The injected content must be clearly wrong given the image
3. Keep the response natural and similar in length
4. The 'text' field must be the EXACT substring in modified_response
5. Return ONLY valid JSON with no markdown fences

Question: <question>
Original response: <original response>

Return this JSON:
{
 "modified_response": "...",
 "injected_span": {"text": "...", "label": "<label>",
 "prob": 1.0}}

Validation

Examine the image carefully.

The following claim was made about this image:
"<injected span text>"

Is this claim FACTUALLY INCORRECT based on what you can see?
Answer with exactly one word:
YES (it is wrong) or NO (it is accurate).

A candidate is retained only if the validation step returns YES; otherwise, it is discarded.

H Synthetic Data Examples

Table 8 illustrates the transformation rules for each grounded heuristic-injection type; examples are constructed, not exact logged instances.

Label	Example
OCR	<i>Seed:</i> "A sign in the image reads 'EXIT'." <i>Synthetic:</i> "A sign in the image reads ' EXIT '."
Miscounting	<i>Seed:</i> "There are three apples on the table." <i>Synthetic:</i> "There are two apples on the table."
Mischaracterization	<i>Seed:</i> "The car is red and parked near a building." <i>Synthetic:</i> "The car is blue and parked near a building."
Other	<i>Seed:</i> "The door appears to be open." <i>Synthetic:</i> "The door appears to be closed ."
Invention	<i>Seed:</i> "A person is standing on the sidewalk." <i>Synthetic:</i> "A person is standing on the sidewalk. There is also a small red bicycle nearby."

Table 8: Illustrative examples of grounded heuristic injection. Bold marks the injected hallucination span.