HALT: HELD-OUT-SOURCE EVALUATION FOR HALLUCINATION DETECTION IN FORENSIC VIDEO REPORTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination detectors for long-form multimodal generation are almost always evaluated on text from the same source LLMs they were trained on. In forensic video reporting we show this conflates detection with identifying which LLM wrote the text. A detector that silently stops working when the underlying model is updated offers false assurance, in a setting where inventing a detail and omitting a crime are two different evidentiary failures. We build HALT (Hallucination Across LLMs and Techniques) on a previously released corpus of 19,361 forensic video reports (21.8M words) in which three frontier multimodal LLMs each describe all 807 surveillance videos, spanning eleven crime categories, under all eight prompting strategies, labelled for six hallucination types by a cross-judging panel in which no model evaluates its own output. Our central result is that much of what a detector appears to know is the identity of the model that wrote the report: a predictor reading no report text at all, returning only the writing model's base rate, captures 70 to 73% of a detector's above-chance signal on the fabrication labels against 13% on omission. Withholding a source LLM accordingly costs the fabrication labels 0.104 to 0.191 AUC against 0.035 to 0.105 for omission and distortion, consistently across four detector families from bag-of-words to a fine-tuned long-context encoder, and again when a source LLM is updated rather than replaced, though the tendency being detected is unchanged across that update. Two further findings concern the panel's H1–H6 labels: (i) report length predicts the panel's scene-fabrication label at AUC 0.727 but the consensus of two human annotators who each independently labelled the same 130 reports at only 0.566, and (ii) models are not allowed to judge their own output, so each source LLM is scored by a different pair of judges; those pairs differ in how often they agree, so comparing hallucination rates across models partly compares their judges. We release the reliability tiers, the four splits, and reference implementations of every detector.

1 INTRODUCTION

Panel-of-judges protocols now label datasets far larger than human annotation could reach, and those labels underpin leaderboards, safety audits and deployment decisions. Whether they measure what they are taken to measure is rarely examined, and the raw verdicts are rarely released.

Hallucination detection inherits both problems. A detector is trained on text labelled this way and tested on text from the same LLMs that produced the training set, split at random. Model families differ sharply in surface style and equally sharply in how often they hallucinate, so a detector with access to both regularities can score well by recognising which family wrote a passage and applying its base rate, without assessing whether any claim is supported. Under a random split this is invisible: the models at test time are the models at training time.

The distinction matters wherever a detector meets a model it has not seen. We study it in forensic video reporting, where inventing a detail and omitting a crime are different evidentiary failures, a screening tool that silently stops working when the model is updated offers false assurance, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

expert reference annotations written independently of any model are available, which most long-form generation tasks lack.

Separating the two requires several model families over identical inputs, so the source LLM can be withheld while everything else is fixed. We ask:

1. Do hallucination detectors transfer to source LLMs they were not trained on, and by how much does performance change?
2. How does detectability depend on hallucination type, and does the type ordering hold under transfer?
3. Which surface properties carry the detectable signal, and what remains once source identity is removed?
4. How much of the loss does conditioning on the reference annotation recover, and for which types?
5. How does panel design shape the labels themselves, both in what drives judge verdicts and in what excluding self-evaluation implies?

The report corpus, the labelling protocol, the human-validation subset and the per-judge verdicts are inherited from Adeniran et al. (2026), which measured how often hallucination occurs and released its materials. Its detectability analysis trained and tested every detector on the same three source LLMs, so it could not ask whether detection transfers, and it names the absence of an out-of-distribution test among its own limitations. New here are: the held-out-source split and three other video-disjoint partitions; reference-conditioned detectors and fine-tuned long-context encoders, so the result cannot be attributed to weak models; per-type reliability tiers; a length-bias diagnosis scoring panel and human labels against the same feature; and a within-judge test of what excluding self-evaluation costs.

We call the LLM that produced a report its *source LLM*, to distinguish it from the LLMs that later judge it. In the language of shortcut learning (Geirhos et al., 2020) and distribution-shift benchmarks (Koh et al., 2021), source identity is a spurious feature: predictive in distribution, cheaply recoverable, worthless once the source changes.

2 Related Work

Existing hallucination benchmarks target short outputs and a fixed set of source LLMs. CHAIR (Rohrbach et al., 2018) measures object hallucination in captions, POPE (Li et al., 2023b) uses polling, HaluEval (Li et al., 2023a) benchmarks 35K samples, THRONE (Kaul et al., 2024) unifies evaluation across six models, Bingo (Cui et al., 2023) identifies bias triggers, and VideoHallucer (Wang et al., 2024c) extends the setting to short clips. Video-specific benchmarks have since proliferated (Li et al., 2025; Rawal et al., 2025; Seth et al., 2025; Yang et al., 2025; Ge et al., 2025; Gao et al., 2025; Wang et al., 2025; Cai et al., 2025). None varies the writing model while holding the input fixed, so none can separate a detector that reads evidence from one that reads style. Detection methods inherit the same limitation: SelfCheckGPT (Manakul et al., 2023), semantic entropy (Kuhn et al., 2023), and FActScore (Min et al., 2023) detect hallucination via consistency, uncertainty, or fact decomposition, but are evaluated within a single source LLM or across several pooled without a held-out condition. Our reference-conditioned baseline is closest to FActScore, simplified to lexical grounding. The closest analogue to our argument lies outside the hallucination literature: The shape is familiar from annotation artifacts, where the hypothesis alone recovers the label two thirds of the time (Gururangan et al., 2018). Wang et al. (2024b) show that detectors of machine-generated text degrade on generators they have not seen, Pu et al. (2023) ask directly how detectors perform on a generator they were not trained on, and RAID (Dugan et al., 2024) makes held-out generators a standard axis of that benchmark. That work asks whether text came from a machine; we ask whether a detector nominally judging content is keyed to the same signal.

Panel designs (Zheng et al., 2023) and cross-model panels that mitigate self-preference (Panickssery et al., 2024) are now standard, and their reliability is typically summarised by a single agreement figure. A growing body of work questions whether that figure means what it appears to: Wang et al. (2024a) document position and scoring biases, Wu & Aji (2023) show judges reward stylistic properties over substance, and Stureborg et al. (2024) find judges inconsistent across prompts and

biased toward text they find predictable. Length bias sits in this family. We report reliability per hallucination type against independent human raters, measure length bias against human labels on the same reports (Section 4.6), and show that the no-self-evaluation constraint itself confounds each source LLM with its judges, which single-figure reporting hides.

The study that produced our raw material (Adeniran et al., 2026) deserves a precise statement of difference, since it supplies the corpus, the taxonomy and the verdicts. It asks how often and in what form frontier MLLMs hallucinate, and treats the panel labels as the ground truth to measure against; we ask whether a detector trained on those labels transfers to a model it has not seen, and treat the labels themselves as an object of study. Its detectability analysis fits logistic regressions and embeddings under five-fold cross-validation with the same three source LLMs on both sides of every split, so the transfer question cannot arise there, and it names the absence of an out-of-distribution test among its own limitations. Withholding a source LLM entirely, tiering labels by reliability, scoring one feature against panel and human labels alike, and testing whether the judging design confounds the source are all outside its scope.

3 METHODOLOGY

3.1 TASK AND CORPUS

An instance is a triple (v, m, τ): a surveillance video v, a source LLM m, and a prompting strategy τ. The source LLM produces a free-form forensic report r describing what occurs in v, and each report carries an expert reference annotation g_v written independently of any model. The task is to predict, from r and optionally g_v, six binary labels: H1 scene fabrication, H5 entity fabrication and H6 phantom actors, which invent content absent from the footage; H3 crime missed, which drops the criminal act; and H2 crime misclassification and H4 severity minimization, which misstate its nature or gravity. These group onto three axes, *fabrication*, *omission* and *distortion*. We report per-type AUC, a macro over the six, a macro restricted to the reliable types (Section 3.2), and a binary any-hallucination score.

The corpus we inherit (Adeniran et al., 2026) is $19{,}361$ reports on 807 UCF-Crime videos (Sultani et al., 2018) with UCA annotations (Yuan et al., 2024) as reference, written by Claude Opus 4.7, GPT-5.5 and Gemini 3.1 Pro (Claude, GPT and Gemini below), the frontier releases of 2026 and all three still current on their providers' APIs at submission, under eight prompting strategies. Two of its properties make the held-out-source question askable. It is fully crossed up to seven failed API calls, so withholding a source LLM holds the input fixed by construction. And verbosity varies by a factor of three under an identical task, 580 mean words per report for GPT against $1{,}802$ for Gemini, the regularity Section 4 shows detectors exploiting.

3.2 LABELS AND THEIR RELIABILITY

The inherited labels come from a panel of the same three models acting as judges (Zheng et al., 2023), each report scored by the two that did not write it, so self-preference bias (Panickssery et al., 2024) is structurally excluded (Adeniran et al., 2026). Three properties of that design matter here. It yields two verdicts per report under a conjunctive rule, so alternative rules can be applied post hoc (Section 4.1). Judge inputs cap the reference at $1{,}500$ and the report at $3{,}000$ characters, which Section 4.6 shows shapes the labels. And no model judging its own output means each source LLM is scored by a distinct judge pair, the confound of Section 4.6.

An inherited subset of 130 reports carries independent labels from two human annotators, members of the research group rather than forensic practitioners, who worked from text with the panel's labels hidden. From the rater-rater and panel-consensus agreements we assign each type a reliability tier and exclude the low tier from the headline macro; the tiering is ours, and Section 4.1 reports it.

3.3 EVALUATION SETUP

Each video contributes 24 reports, so a split over reports would let a detector memorise the video rather than learn the phenomenon; Figure 1 shows the three partitions. The standard and held-out-strategy splits are therefore video-disjoint. The held-out-source split is not, since every video is

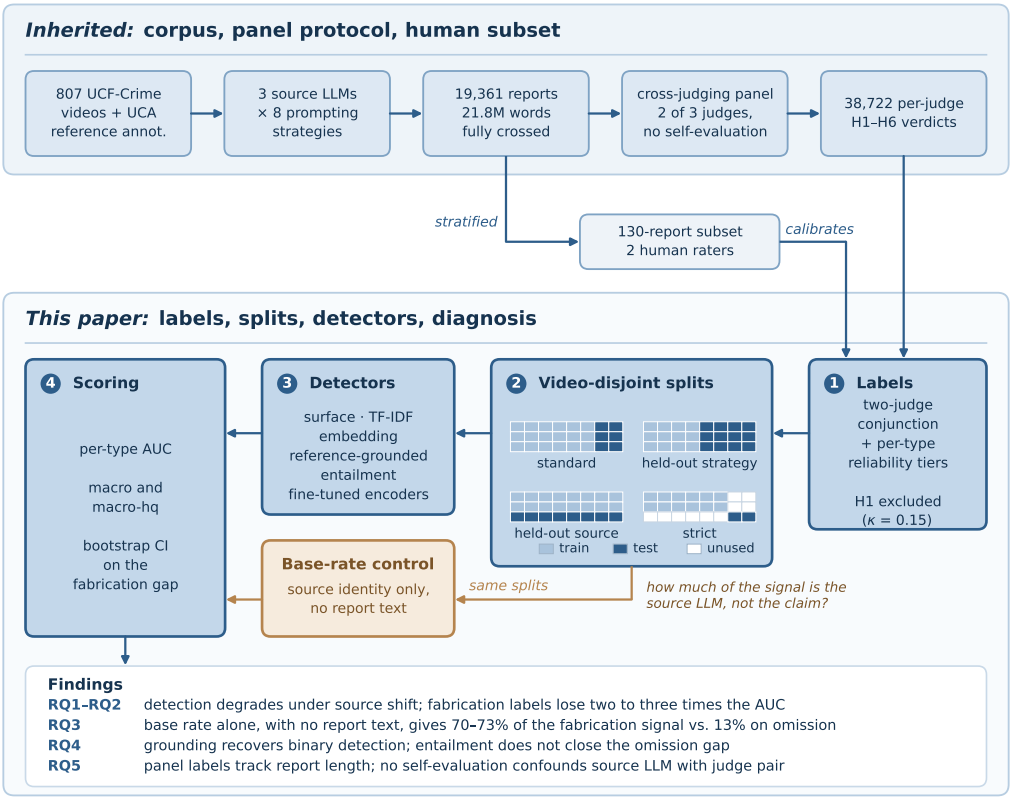


Figure 1: The HALT pipeline. The shaded band is inherited; the remainder is this paper: aggregating the raw verdicts into labels whose per-type reliability is calibrated against an independently scored subset, partitioning the corpus into four video-disjoint splits, and evaluating four detector families against a base-rate control that reads no report text. In step 2 each cell is one report, rows are source LLMs and columns are video-by-strategy conditions; existing benchmarks admit only the standard split.

described by every source LLM; Section 4.3 reports a disjointness-enforced variant under which the effect is larger, so the main tables are conservative.

We provide reference detectors in two families. The first is linear: five logistic regressions with balanced class weights, differing only in features. Hand-built features describing how a report is written, rather than what it claims, make up **surface**: verbosity, sentence statistics, hedging density, numeric detail density, type-token ratio, truncation length, and source and strategy identity indicators. **surface-noid** is the same set without the identity indicators; **tfidf** uses word 1–2 gram TF-IDF; **embed** uses sentence embeddings; and **grounded** uses report-versus-reference features only, namely the rate of report tokens absent from the reference, the rate of reference tokens absent from the report, Jaccard overlap, reference coverage, length ratio and TF-IDF cosine. Identity features are dropped automatically on any split where they are constant in training and unseen at test.

The second family is fine-tuned encoders. One encoder per split, six per-type heads plus a binary head, report and reference as a sentence pair, positive weights matching the linear baselines. Context length is chosen rather than defaulted: the judge saw about 1,125 tokens, so a detector at 512 sees less than the judge did. We report **DeBERTa-v3-base** at its 512-token limit, **ModernBERT-base** at 1,280, matching the judge's view, and ModernBERT at 2,048, exceeding it.

The corpus, per-judge verdicts, aggregated labels and rubrics were released with the prior study. HALT adds, and releases: the reliability tiers with supporting n, the four splits, reference implementations of every detector, and a datasheet (Gebru et al., 2021). The artifact is available at

https://anonymous.4open.science/r/HALT-ANON-7FF2/. Supporting tables cited below are in Appendices A to H of this paper.

4 RESULTS

4.1 CORPUS STATISTICS AND LABEL PROPERTIES

The prior study reports 91.1% of reports carrying at least one hallucination under the released rule. What it does not report is how far that depends on the rule: a disjunctive rule on the same verdicts gives 98.5% and 3.55 per report, a seven-point swing that a release of aggregated labels alone would conceal.

Reliability is strongly type-dependent (Table 1): κ against the human consensus runs from 0.81 for entity fabrication down to 0.15 for scene fabrication, which we exclude from the headline macro. Two qualifications apply. The consensus κ is computed only where the two raters agree, a subset of 62 to 110 of 130, so H5's 0.81 rests on 64 reports and, since the restriction selects the straightforward cases, κ_{panel} is an upper bound. The raters also agree only modestly with each other on fabrication (macro $\kappa = 0.28$, $\kappa \approx 0$ on H1), and the disagreement has a direction: because references are summary-level, granular detail absent from the summary can be scored as fabrication. Absolute fabrication prevalence is therefore an upper bound; relative comparisons are unaffected, since the bias is shared across source LLMs. Because judge inputs truncate the report at $3{,}000$ characters and the most verbose source LLMs are the most truncated, truncation can only understate their fabrication rates.

4.2 SAME-SOURCE DETECTION REPRODUCES PRIOR FINDINGS

Under the standard video-disjoint random split we reproduce the prior study's detectability finding, including that source identity is its most informative single feature: surface features reach 0.858 on scene fabrication against 0.646 on crime omission, and TF-IDF reaches 0.846 macro and 0.820 binary. That is what a practitioner would conclude under conventional evaluation.

4.3 DETECTION OF THE FABRICATION LABELS DEGRADES MOST (RQ1–RQ2)

Every detector degrades under leave-one-source-out, and every one degrades unevenly (Table 2, Figure 2): fabrication falls by 0.104 to 0.191 AUC against 0.035 to 0.105 for omission and distortion, a ratio of 1.4 to 3.1. We describe this as degradation of the fabrication *labels*: the raters agree least on exactly these types (Section 4.1), so a property of fabrication and a property of how fabrication is labelled here cannot be separated. For the lexical detector the loss inverts the per-type ordering

Table 1: Per-type label reliability on the 130-report human-annotated subset. The agreement figures are from the prior study (Adeniran et al., 2026); the tiering and its use to gate the headline macro are ours. κ_{panel} compares the panel against the human consensus and is computed only on reports where the two raters agree, so it is an upper bound; n is that subset size. Thresholds are $\kappa \geq 0.65$ (high) and $\kappa \geq 0.40$ (medium). We tier on κ_{panel} and exclude only H1 from the headline macro. The last column shows what the tier would be on rater-rater agreement instead: the two entity-level fabrication types would fall to low, which would leave no fabrication type in the macro at all. Our fabrication results should be read against that.

Type	Axis	n	$\kappa_{\text{human-human}}$	κ_{panel}	Tier (panel)	Tier (raters)
H1 Scene Fabrication	Fabrication	62	-0.008	0.150	low	low
H2 Crime Misclassification	Distortion	99	0.470	0.700	high	medium
H3 Crime Missed	Omission	110	0.532	0.432	medium	medium
H4 Severity Minimization	Distortion	103	0.406	0.583	medium	medium
H5 Entity Fabrication	Fabrication	64	0.082	0.811	high	low
H6 Phantom Actors	Fabrication	84	0.192	0.658	high	low
Macro			0.279	0.556		

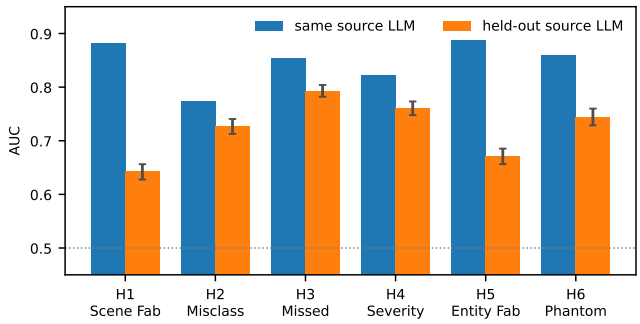


Figure 2: Per-type detector AUC (TF-IDF features) under the standard same-source split and under leave-one-source-out, with bootstrap 95% intervals on the held-out bars. The fabrication types (H1, H5) lose roughly four times as much as the omission and distortion types, inverting the ordering a same-source evaluation would report.

Table 2: Degradation from the standard split to leave-one-source-out, by axis, for four detector families. Drops are LOSO minus standard. The ratio is the fabrication drop divided by the larger of the omission and distortion drops, the conservative choice; it is quoted in this form throughout. Fabrication degrades most in every trained family. The prompted detector is listed for contrast: it is untrained on these labels and emits hard $0/1$ verdicts rather than scores, so it is not directly comparable to the rows above. Its binary column is omitted because it flags some hallucination in most reports, leaving that outcome near-constant.

	drop by axis				ANY		
Detector	fabr.	omis.	dist.	MACRO	standard	LOSO	ratio
tfidf (linear)	−0.191	−0.061	−0.054	−0.123	0.820	0.566	3.1×
DeBERTa-v3, 512	−0.104	−0.035	−0.043	−0.072	0.861	0.729	2.4×
ModernBERT, 1,280	−0.163	−0.089	−0.097	−0.129	0.833	0.644	1.7×
ModernBERT, 2,048	−0.145	−0.079	−0.105	−0.120	0.832	0.642	1.4×
Qwen3.8-27B, prompted	−0.089	−0.005	−0.020	−0.052	—	—	4.5×

outright, TF-IDF giving fabrication 0.686 against omission 0.793 under shift having given 0.876 against 0.854 in distribution.

The difference is statistically unambiguous. Bootstrapping the fabrication degradation minus the omission-and-distortion degradation over 1,000 resamples of both splits gives -0.134 with a 95% interval of $[-0.148, -0.121]$, excluding zero. Per type (Table 8): bootstrap 95% intervals place crime omission above 0.776 in every fold, while scene fabrication never exceeds 0.714 and entity fabrication never 0.730, and the intervals do not overlap in any fold.

The effect is not an artifact of residual video overlap: enforcing video disjointness as well lowers every number and widens the gap, fabrication to 0.668 against omission 0.780 and binary detection to 0.559 (Appendix D).

For the lexical detector binary detection all but disappears: any-hallucination AUC falls from 0.820 to 0.566, and on the held-out Claude fold the interval is $[0.502, 0.578]$, indistinguishable from a coin flip.

A detector that cannot learn the shortcut shows the asymmetry anyway. Prompting Qwen3.8-27B zero-shot under the same splits, a model from outside the three that wrote and judged this corpus and so with no source-specific base rate available, still loses 0.089 AUC on fabrication against 0.005 on omission and 0.020 on distortion. That is evidence against reading the asymmetry as purely a shortcut artifact, and consistent with the fabrication labels being the harder thing to predict on an unseen source. It emits hard verdicts rather than scores, and flags some hallucination in most reports, so we omit its binary column.

Table 3: How much of a detector's above-chance signal source identity alone supplies. The base-rate predictor sees no report text: it returns the training-set prevalence of its label for the source LLM that wrote the report. Share is $(\text{base-rate} - 0.5)/(\text{tfidf} - 0.5)$, both on the standard video-disjoint split.

Type	Axis	base rate	tfidf	share
H1 Scene Fabrication	fabrication	0.777	0.882	73%
H5 Entity Fabrication	fabrication	0.773	0.888	70%
H6 Phantom Actors	fabrication	0.662	0.859	45%
H2 Crime Misclassification	distortion	0.595	0.773	35%
H4 Severity Minimization	distortion	0.599	0.822	31%
H3 Crime Missed	omission	0.545	0.854	13%
ANY	binary	0.725	0.820	70%

The same holds for a version update, which is the narrower failure a deployment meets. Training on current-generation reports and testing on a previous-generation pilot of the same videos (Appendix H) drops binary detection from 0.820 to 0.609, against 0.566 when a family is withheld, fabrication again losing most. The behaviour being detected does not change across that gap: each model's dominant axis is identical in both generations. A detector reading a source LLM's standing tendency would survive the update; this one does not, which suggests it had learned the surface form of a version rather than the tendency it accompanied.

Nor is it an artifact of the low-reliability type. Dropping H1 from the fabrication axis, as the headline macro already does, leaves the ordering intact: the ratio falls to 2.7, 2.2, 1.5 and 1.1 respectively. Entity fabrication and phantom actors alone carry the effect.

The encoders retain far more under shift, with binary detection at 0.642 to 0.729 against 0.566, and fabrication remains the worst-hit axis for every one. The size of the asymmetry does not track detector strength: DeBERTa at 512 tokens is the strongest detector under shift on both binary AUC and macro loss, yet its ratio of 2.4 exceeds either ModernBERT's. Within the encoder family it falls as context grows, 2.4 to 1.7 to 1.4, but context does not explain it either, since TF-IDF sees the whole report untruncated and has the largest ratio of all. We cannot say whether the ratio plateaus above one. For screening at volume, the cheapest encoder is thus also the most robust, DeBERTa processing 2.5 times fewer tokens per report than ModernBERT at 1,280.

Context is not the binding constraint. ModernBERT at 1,280 tokens sees what the judge saw, at 2,048 it sees more, and the two are indistinguishable under shift (binary 0.644 against 0.642), consistent with the labels rather than the detector's input being the limiting factor, though it does not establish that.

Removing the identity features does not help: style is recoverable from the text itself (Appendix B). How much of the signal source identity accounts for is measured next.

4.4 Source identity is recoverable, and carries most of the signal (RQ3)

The shortcut can be measured rather than inferred. A TF-IDF classifier predicting which of the three source LLMs wrote a report reaches 1.000 accuracy on held-out videos against a chance rate of 0.333.

What that buys depends sharply on the type (Table 3). A predictor with no access to the report, returning only the source LLM's training-set prevalence, reaches 0.777 on scene fabrication and 0.773 on entity fabrication against the lexical detector's 0.882 and 0.888. On the above-chance scale that is 73% and 70% of what the detector achieves. On crime omission it reaches 0.545 against 0.854, or 13%. Detection of the fabrication labels is therefore mostly source recognition and omission is mostly not, which is the mechanism behind the transfer asymmetry.

4.5 What grounding buys, and what it does not (RQ4)

Among the linear detectors, reference-conditioned features are the only ones that move binary detection under shift, reaching 0.702 against 0.566 for TF-IDF (Appendix G), and they give that family's

Table 4: Does report length predict the panel label or the human label? AUC of a single feature, report word count, against each label source on the 130-report subset. The human column is restricted to reports where both raters agree (n from Table 1). A large positive gap means the feature tracks the judge rather than the phenomenon. Other features are in Appendix F.

Type	AUC vs. panel	AUC vs. human	Gap
H1 Scene Fabrication	0.727	0.566	$+\mathbf{0.161}$
H4 Severity Minimization	0.505	0.275	$+\mathbf{0.230}$
H3 Crime Missed	0.484	0.390	$+0.093$
H2 Crime Misclassification	0.621	0.608	$+0.013$
H5 Entity Fabrication	0.784	0.800	-0.016
H6 Phantom Actors	0.662	0.668	-0.006

best fabrication axis. The encoders exceed that from the raw reference alone. They receive the report and the reference as a sentence pair and learn their own comparison, with none of the engineered overlap features; what those features buy the linear family, capacity buys the encoder family, and neither removes the asymmetry of Table 2.

One surface cell is diagnostic: on the held-out-Gemini fold severity minimization falls to 0.443, below chance, so the relationship learned in distribution reverses on an unseen source rather than merely weakening, as a shortcut does when its feature is anti-correlated in the held-out family.

Grounding is, however, worse on omission: 0.577 against TF-IDF's 0.793. Our grounded features are lexical by construction, so we tested a semantic version: a cross-encoder NLI model scores every report sentence against the reference and, in the reverse direction, every reference sentence against the report, aggregated into report-level entailment and contradiction features. The reverse direction is what should capture a dropped criminal act. It helps omission and does not rescue it, 0.577 to 0.625 under shift, while costing elsewhere: the entailment detector reaches 0.632 binary against grounding's 0.702 and 0.648 on omission in distribution against TF-IDF's 0.854. Its degradation ratio, 4.8, is the largest of any detector we ran. Whether a report drops the criminal event is therefore not a lexical-representation problem, and claim-decomposed matching (Min et al., 2023) may fare no better. No detector wins everywhere, and binary detection under shift remains the benchmark's open problem.

4.6 Two properties of the labels themselves (RQ5)

Two properties of the labels limit what they can support. The first is a length bias, exposed by scoring a single feature against both the panel and the human labels on the shared 130 reports (Table 4). Report length predicts the panel's scene-fabrication label at AUC 0.727 but the human label at only 0.566, a gap of 0.161. Severity minimization diverges more sharply still: for humans, length runs inverse to the label (0.275), while the panel sees no relationship (0.505). Entity fabrication and phantom actors show no such gap (0.784 versus 0.800, 0.662 versus 0.668), so the bias is specific rather than pervasive: it says which labels to distrust, not that all are suspect.

Part of the effect is mechanical. Judge inputs cap at 3,000 characters: the truncation flag alone predicts the panel's scene-fabrication label at 0.686, and within truncated reports length falls to 0.537. It does not dissolve there: on the 45 untruncated reports the panel reaches 0.669 against the human's 0.566, consistent with the labelling mechanism of Section 4.1, where granular detail absent from a summary-level reference is scored as fabrication. The bias is real but smaller than the pooled 0.727 suggests.

The second property is structural. Excluding self-evaluation has an unavoidable consequence: each source LLM is scored by a different pair of judges. Those pairs differ sharply in mutual agreement. Cohen's κ on scene fabrication is 0.850, 0.415 and 0.148 for the pairs scoring Claude, GPT and Gemini respectively (Appendix E), so under a conjunctive rule a disagreeing pair yields fewer positives for purely mechanical reasons. The pairwise agreements are published with the corpus; what has not been drawn from them is that under no-self-evaluation they attach to source LLMs rather than to judges. That makes them a confound between the source LLM and its judges rather than a reliability statistic, and it follows from the design rather than from this corpus.

Table 5: Judge-pair agreement and within-judge axis rates. The pairs differ by a factor of five on the fabrication types; the within-judge rates hold the judge fixed, and the dominant axis is reproduced by every judge for every source LLM it sees.

(a) Pair agreement, Cohen's κ

Source LLM	Judged by	H1	H5	overall
Claude	GPT + Gemini	0.850	0.845	0.809
GPT	Claude + Gemini	0.415	0.475	0.580
Gemini	Claude + GPT	0.148	0.138	0.614

The confound is testable: each judge scores exactly two source LLMs, so a within-judge comparison holds the judge fixed (Appendix E). Across 18 judge-by-type-by-pair comparisons the ordering given by the aggregated labels is reproduced in 16, and at the axis level in all six: every judge that sees GPT calls it omission-dominant, and every judge that sees Claude or Gemini calls them fabrication-dominant.

5 Limitations

Three source LLMs. Three folds is thin evidence for a claim about model families in general. Newer releases have appeared since, though the held-out-generation test indicates the effect is not generation-specific.

Fabrication labels, not fabrication. Our raters agree least on these types (macro $\kappa = 0.28$) and the panel's label tracks report length where theirs does not, so a property of fabrication and a property of how it is labelled here cannot be separated. The asymmetry survives dropping the weakest type, and appears in a prompted detector with no source base rate to learn, which points at the labels rather than the shortcut alone.

One domain. The magnitudes should not be assumed to carry to summarisation, captioning or dialogue; settling that needs a second corpus crossed over generators.

6 Conclusion

Hallucination detectors are evaluated almost exclusively on text from the source LLMs they were trained on, which cannot separate detecting the phenomenon from recognising the source. Here the source is trivially recoverable, and withholding one costs the fabrication labels several times what it costs omission, across four detector families and again when a source is updated rather than replaced. The magnitudes are specific to this domain; the evaluation gap that hid them is not.

Ethics Statement

The corpus derives from a public surveillance benchmark of de-identified footage; annotators worked from text only. We release it for auditing and detector development, not for automated legal or investigative decision-making: a system whose error rate depends on the source LLM, the prompting strategy and the event type does not meet the standard evidentiary use requires.

AI Use Statement

An AI assistant was used to draft and revise sections of this paper from the authors' direction.

Reproducibility Statement

A single entry point rebuilds every label from the released per-judge verdicts and regenerates every table and figure, verifying the published corpus statistics on each run. All splits are seeded; Section 3.3 lists the artifacts and the repository link.

REFERENCES

Opeyemi Adeniran, Jamell Dacon, Derrick Cook, Temitope Ajibola, Kelechi Nwachukwu, Peter Taiwo, and Kofi Nyarko. Beyond accuracy: A type-aware taxonomy and evaluation framework for prompted hallucination in forensic multimodal llms. In *IEEE International Conference on Data Mining (ICDM)*, 2026. To appear.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Jiahao Chen, Xiaotao Gu, and Minlie Huang. MHALO: Evaluating MLLMs as fine-grained hallucination detectors. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Lina Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *arXiv preprint arXiv:2503.19622*, 2025.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Nuo Chen, Junying Chen, Zhihong Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, and Benyou Wang. MLLM-Bench: Evaluating multimodal LLMs with per-sample criteria. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 107–112, 2018.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733, 2025.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. On the zero-shot generalization of machine-generated text detectors. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4799–4808, 2023.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and omission evaluation in Video-LLMs. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20280–20290, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. EgoIllusion: Benchmarking hallucinations in egocentric video understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 28461–28480, 2025.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9440–9450, 2024a.

Sheng Wang, Yi Zhang, Yang Zhu, Enhao Liu, Jian Li, Yang Liu, and Xiangyang Ji. SHIFT: Smoothing hallucinations by information flow tuning for multimodal large language models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3639–3649, 2025.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024b.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024c.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*, 2023.

Shuo Yang, Wei Yu, Wen Yang, Xin Liu, Hao Tan, Long Lan, and Nong Xiao. WildVideo: Benchmarking LMMs for understanding video-language interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22052–22061, 2024.

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. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

APPENDIX

The following material supplements the main text and is referenced from it. Tables are numbered continuously with the main text.

A HELD-OUT-STRATEGY RESULTS

Table 6: Held-out-strategy split: train on the four single-turn strategies, test on the four multi-turn ones. Compare the macro column against the standard split (Table 7) and against leave-one-source-out (Table 2).

Features	H1	H2	H3	H4	H5	H6	ANY	MACRO	MACRO-HQ
surface	0.809	0.601	0.547	0.500	0.800	0.736	0.732	0.665	0.637
surface-noid	0.794	0.593	0.555	0.534	0.790	0.719	0.703	0.664	0.638
tfidf	0.868	0.753	0.859	0.832	0.873	0.839	0.808	0.837	0.831
embed	0.745	0.681	0.617	0.730	0.632	0.624	0.633	0.671	0.657
grounded	0.837	0.690	0.602	0.358	0.821	0.769	0.746	0.679	0.648
DeBERTa-v3, 512	0.842	0.816	0.871	0.879	0.870	0.876	0.810	0.859	0.862
ModernBERT, 1,280	0.826	0.777	0.844	0.863	0.857	0.873	0.787	0.840	0.843
ModernBERT, 2,048	0.828	0.786	0.853	0.873	0.854	0.871	0.785	0.844	0.847

B SAME-SOURCE BASELINES AND THE IDENTITY ABLATION

Table 7: Baseline AUC by split for the surface feature families. The held-out-source rows are the held-out-Gemini fold, not the three-fold mean; Table 8 gives the per-fold values and means for TF-IDF. Removing the source and strategy identity indicators (surface-noid) does not remove the shortcut, because style is recoverable from the text itself: scene fabrication still reaches 0.811 under the standard split and still collapses to 0.603 on this fold.

Features	Split	H1	H2	H3	H4	H5	H6	ANY	MACRO-HQ
surface	random	0.858	0.690	0.646	0.653	0.865	0.772	0.765	0.725
	held-out strategy	0.809	0.601	0.547	0.500	0.800	0.736	0.732	0.637
	held-out src (Gemini)	0.628	0.646	0.506	0.443	0.559	0.564	0.583	0.544
surface-noid	random	0.811	0.667	0.576	0.636	0.830	0.738	0.720	0.689
	held-out strategy	0.794	0.593	0.555	0.534	0.790	0.719	0.703	0.638
	held-out src (Gemini)	0.603	0.614	0.470	0.538	0.567	0.632	0.582	0.564
tfidf	random	0.882	0.773	0.854	0.822	0.888	0.859	0.820	0.839
	held-out strategy	0.868	0.753	0.859	0.832	0.873	0.839	0.808	0.831
	held-out src (Gemini)	0.630	0.727	0.795	0.715	0.621	0.720	0.613	0.716

C PER-FOLD LEAVE-ONE-SOURCE-OUT

Table 8: TF-IDF features, per fold, with bootstrap 95% confidence intervals (1,000 resamples). The crime-omission interval overlaps neither the scene- nor the entity-fabrication interval in any fold, which is the separation cited in Section 4.3.

Held-out source LLM	H1	H2	H3	H4	H5	H6	ANY
Claude	0.596	0.718	0.787	0.788	0.671	0.778	0.539
	[.581,.611]	[.704,.732]	[.776,.799]	[.774,.801]	[.658,.686]	[.765,.788]	[.502,.578]
GPT	0.702	0.735	0.796	0.779	0.719	0.736	0.548
	[.689,.714]	[.719,.749]	[.785,.807]	[.767,.790]	[.707,.730]	[.713,.760]	[.531,.566]
Gemini	0.630	0.727	0.795	0.715	0.621	0.720	0.613
	[.613,.644]	[.715,.741]	[.785,.806]	[.702,.729]	[.604,.640]	[.708,.732]	[.585,.636]
Mean (held-out)	0.643	0.727	0.793	0.761	0.671	0.744	0.566
Random split	0.882	0.773	0.854	0.822	0.888	0.859	0.820
Degradation	−0.239	−0.046	−0.061	−0.061	−0.217	−0.115	−0.254

Table 9: AUC on the held-out source LLM, per fold, for every feature set. Holding out GPT, the only omission-dominant family, drops embedding scene-fabrication AUC to 0.527 while raising crime-omission AUC to 0.824; holding out Claude, leaving the similarly fabrication-dominant Gemini, raises scene fabrication to 0.853.

Features	Held out	H1	H2	H3	H4	H5	H6	ANY
embed	Claude	0.853	0.659	0.689	0.674	0.862	0.661	0.670
	GPT	0.527	0.624	0.824	0.716	0.566	0.673	0.587
	Gemini	0.671	0.627	0.645	0.644	0.655	0.560	0.528
grounded	Claude	0.883	0.612	0.572	0.546	0.840	0.566	0.852
	GPT	0.665	0.596	0.642	0.528	0.639	0.579	0.616
	Gemini	0.649	0.720	0.518	0.523	0.616	0.673	0.638
DeBERTa-v3, 512	Claude	0.934	0.811	0.863	0.847	0.928	0.782	0.816
	GPT	0.708	0.845	0.913	0.864	0.746	0.748	0.665
	Gemini	0.630	0.845	0.889	0.888	0.692	0.818	0.707
ModernBERT, 1,280	Claude	0.885	0.656	0.785	0.731	0.894	0.756	0.739
	GPT	0.548	0.721	0.833	0.812	0.622	0.669	0.592
	Gemini	0.580	0.692	0.707	0.737	0.631	0.777	0.601
ModernBERT, 2,048	Claude	0.877	0.684	0.803	0.750	0.897	0.792	0.738
	GPT	0.602	0.708	0.807	0.783	0.664	0.721	0.601
	Gemini	0.576	0.744	0.798	0.786	0.644	0.787	0.588

D Video-Disjoint Leave-One-Source-Out

Holding out a source LLM leaves the same videos on both sides of the split, since every video is described by every source LLM. Re-running the three folds with video disjointness additionally enforced (training on 9,003 reports, testing on 1,016) lowers every number and widens the gap: the fabrication axis falls to 0.668 and omission to 0.780, so the omission-minus-fabrication margin grows from 0.107 to 0.112, and binary detection falls to 0.559. The figures in the main text are therefore the conservative ones.

E The Judge-Pair Confound

The pair agreements and within-judge rates are in Table 5 of the main text.

F Label-Source Divergence, All Features

Table 10: AUC of each single feature against the panel label and against the human label, on the 130-report subset. Table 4 reports the report-length column. Every reference-conditioned feature diverges in the opposite direction: on grounded quantities the panel does not diverge from the human raters.

Type	novel rate		missing rate		hedging rate	
	panel	human	panel	human	panel	human
H1 Scene Fabrication	0.535	0.546	0.300	0.536	0.551	0.451
H2 Crime Misclassification	0.545	0.600	0.500	0.565	0.471	0.573
H3 Crime Missed	0.618	0.644	0.671	0.692	0.522	0.547
H4 Severity Minimization	0.547	0.661	0.608	0.730	0.616	0.636
H5 Entity Fabrication	0.467	0.595	0.254	0.342	0.464	0.577
H6 Phantom Actors	0.567	0.607	0.408	0.450	0.461	0.450

G Feature Sets Under Leave-One-Source-Out

Table 11: Feature sets and encoders under leave-one-source-out, averaged over the three folds. Axis scores are the mean per-type AUC within the axis. Among the linear detectors, reference-conditioned features are the only ones that recover binary detection; the fine-tuned encoders exceed them without grounding.

Features	Fabrication	Omission	Distortion	MACRO	ANY
surface	0.584	0.506	0.545	0.558	0.583
nli-grounded	0.599	0.625	0.526	0.579	0.632
embed	0.670	0.719	0.657	0.674	0.595
tfidf	0.686	0.793	0.744	0.723	0.566
grounded	0.679	0.577	0.587	0.631	0.702
grounded+surface	0.691	0.607	0.599	0.646	0.652
DeBERTa-v3, 512	**0.776**	**0.888**	**0.850**	**0.820**	**0.729**
ModernBERT, 1,280	0.707	0.775	0.725	0.724	0.644
ModernBERT, 2,048	0.729	0.803	0.742	0.746	0.642
Qwen3.8-27B, prompted	0.677	0.839	0.820	0.779	—

H Rubrics and Full Corpus Statistics

The verbatim judge rubric, the human-rater instructions, the per-cell hallucination rates for all 24 source-LLM-by-strategy combinations, and the datasheet are in the released artifact, which also regenerates every table above from the raw per-judge verdicts.

Both failures fall on the GPT judge and on fabrication types, where it flags nearly everything it sees (92.0% and 96.1% on scene fabrication) and cannot discriminate. The Claude-versus-Gemini fabrication ordering rests on a 4 to 7 point within-judge margin and is judge-dependent; the two orderings involving GPT, at 29 to 41 points and reproduced by every judge, are robust.

Table 12: Within-judge axis rates (%). Each judge scores exactly two source LLMs, so holding the judge fixed isolates the source.

(b) Within-judge axis rates (%)

Judge	Source LLM	Fabrication	Omission	Distortion	Dominant
Claude	GPT	28.6	55.7	42.7	omission
Claude	Gemini	66.7	54.3	39.8	fabrication
GPT	Claude	77.8	41.7	36.0	fabrication
GPT	Gemini	79.2	49.4	38.1	fabrication
Gemini	Claude	72.7	44.5	26.8	fabrication
Gemini	GPT	44.1	53.1	33.1	omission

H Held-out strategy and held-out generation

The held-out-strategy split isolates what is specific to the source LLM rather than to distribution shift in general. Training on single-turn strategies and testing on multi-turn ones costs TF-IDF almost nothing: macro 0.846 to 0.837, binary 0.820 to 0.808 (Appendix A), and the other feature sets behave the same way. A detector that has never seen multi-turn reports handles them; one that has never seen a source LLM does not.

Transfer also depends on whether a similar failure profile appears in training: holding out GPT, the only omission-dominant family, drops embedding scene-fabrication AUC to 0.527 (Table 9), while holding out Claude raises it to 0.853.

The divergence is specific to verbosity, not to disagreement in general. The same comparison for the reference-conditioned features gives negative gaps on almost every type, reaching -0.236 on scene fabrication. Only report length predicts the judge better than it predicts a human reader.

HELD-OUT GENERATION

Leave-one-source-out withholds a model family. A deployment meets a narrower change, a version update within one family, which the corpus cannot pose because every report comes from one generation. A pilot reported by Adeniran et al. (2026) ran the same eight-technique protocol on the previous generation of each family (Claude Sonnet 4, GPT-4o, Gemini 2.0 Flash) over 19 videos, 15 of which are also in our corpus.

We train TF-IDF on current-generation reports of the other 792 videos and test on the 456-report pilot restricted to those 15 videos, 360 reports. The split is video-disjoint, so the task, the videos and the label scheme are fixed and only the generation changes. Pilot labels are rebuilt from the raw per-judge verdicts under the same conjunctive rule as the released labels.

Table 13: Held-out generation, TF-IDF. Intervals are 1,000-resample bootstrap percentiles on the 360-report test set.

Target	AUC	95% CI	positive rate
H1 scene fabrication	0.688	[0.599, 0.780]	0.100
H2 crime misclass.	0.820	[0.766, 0.872]	0.136
H3 crime missed	0.791	[0.731, 0.843]	0.244
H4 severity minim.	0.706	[0.488, 0.894]	0.025
H5 entity fabrication	0.736	[0.670, 0.797]	0.217
H6 phantom actors	0.689	[0.599, 0.768]	0.131
binary	0.609	[0.550, 0.666]	0.533

By axis: fabrication 0.704, omission 0.791, distortion 0.763, against 0.876, 0.854 and 0.797 in distribution. The ordering inverts as it does under held-out source, where the same detector gives 0.686, 0.793 and 0.744.

Two limits. At 360 reports the fabrication and omission intervals overlap, which they do not in any fold of the main corpus, so the separation is directional here rather than established. And the judging models differ across generations as well as the generators, so this measures generator-and-judge shift together; the prior study says the same of its own cross-generation check, and compares only profile shape for that reason. We therefore report this as corroboration of the held-out-source result rather than as an independent confirmation of it.