VERIFIABILITY, NOT CONFIDENCE: CHEAP MULTIMODAL RUMOR DETECTION THAT KNOWS WHEN TO REFUSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal misinformation detection faces a practical trade-off, in which discriminative detectors are cheap but less accurate while agentic pipelines are accurate but costly. We propose a minimal 4B vision–language classifier that answers each post in one forward pass, without an agent or retrieval. It approaches recently published detectors on AMG, FineFake, and Weibo at a measured 1.7×10^{-5} USD per post. Auditing the residual errors reveals a deeper failure: many posts assert no verifiable claim, yet on content the detector never models its confidence is at its highest. We call this property *verifiability* and show that, under a fixed judge rubric, it is linearly decodable from frozen MLLM representations, with AUROC $0.95 / 0.98 / 0.96$ on AMG (in-domain), FineFake, and Weibo (both zero-shot transfers), and that the same readout transfers from English to Chinese and across model families down to 0.9B parameters while remaining near-orthogonal to confidence. A verifiability gate restores rejection where confidence fails, and a cheap resolver behind it keeps the cascade at or above direct accuracy. We believe verifiability offers a principled coordinate for abstention in multimodal detection.

1 INTRODUCTION

Multimodal misinformation, that is, misleading text paired with a supporting or manipulated image, circulates widely across social platforms. A fabricated claim beneath a real photo is more persuasive than either modality alone: a genuine protest photograph captioned with an invented death toll reads as evidence of the toll. Every such post must be judged before it spreads. This motivates *multimodal rumor detection* (Wang et al., 2018): given a text–image post, predict whether it reports real (0) or rumor (1) information.

Existing approaches sit at two ends of a cost–accuracy spectrum. Discriminative detectors answer in a single forward pass but sit below the state of the art; early-fusion and cross-attention models reach 72–76% on AMG, where the best systems reach 80% (Table 19), while agentic pipelines close that gap only by spending 5–80 online LLM calls and \$0.02–1.64 per post. Accuracy is thus either not cheap enough or not accurate enough to deploy. Two questions therefore arise. First, how much accuracy does a cheap, single-forward detector actually leave on the table? Second, when it errs, can we tell which posts it should not answer at all?

We answer the first question with a minimal setup: a 4B vision–language model used as a plain classifier (frozen backbone, LoRA (Hu et al., 2022), a linear head, one forward pass, no agent or retrieval). It already approaches recently published specialized detectors on AMG, FineFake, and Weibo ($0.862/0.809/0.961$) at a measured 1.7×10^{-5} USD per post (GPU-rental accounting, Appendix Q).

Auditing the errors of this detector answers the second question, and the answer is a phenomenon rather than a further accuracy gap. Most residual errors are not misjudged claims: they sit on posts asserting no proposition checkable against public evidence, such as opinions, jokes, and life sharing, where any real/rumor output is ungroundable and no detection capacity settles them. Confidence cannot see these posts, because it only measures difficulty inside the trained label space: the margin ranks the detector's own errors there (AUROC 0.735 on AMG), yet on MR2 posts of a class the detector never models carry far higher margins than ordinary posts, and a confidence gate refuses none

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.

of them (Section 3). We call the property that separates these posts *verifiability*, namely whether the post asserts a verifiable claim at all, and we operationalize it as a human-audited LLM-judge verdict under a fixed rubric.

That makes it something the detector can read rather than guess: because verifiability is settled by the post and not by the model's knowledge, a linear probe on frozen hidden states recovers it at AUROC $0.95 / 0.98 / 0.96$ with no external knowledge and no retrieval, and the same readout transfers zero-shot from English to Chinese and across model families down to 0.9B parameters while remaining near-orthogonal to confidence. Turned into a rule, the readout refuses the posts the detector cannot ground, and a second-level handler behind it answers every refusal. The detector is therefore selective, but the full system never is.

Our contributions are threefold:

- In Section 3 we identify a failure mode that the real/rumor objective leaves unmodeled: a cheap single-forward detector's residual errors concentrate on posts that assert no verifiable claim, and confidence is structurally blind to them.
- In Section 4 we turn that phenomenon into the rejection layer these systems lack: a linear readout of verifiability, a two-gate rule whose direction is set automatically, and a cascade that answers what the gate refuses.
- In Section 5 we measure it end to end: the gate restores rejection where confidence refuses nothing, the cascade holds fully-answered accuracy above the seed-matched detector on all three datasets at $18\times$–$1500\times$ lower cost than the agentic pipelines of Table 19, and a controlled negative result shows that the direction marks the failure rather than causing it.

2 Related Work

2.1 Detectors: Accuracy Bought with Cost

Detector architectures have progressed through three generations, each buying accuracy with cost. Early fusion models concatenate per-modality features: EANN (Wang et al., 2018), SpotFake (Singhal et al., 2019), MVAE (Khattar et al., 2019). Cross-modal attention learns text–image interactions: MCAN (Wu et al., 2021), CAFE (Chen et al., 2022). A third line targets *out-of-context* misinformation, where veracious text and image are paired inconsistently: NewsCLIPpings (Luo et al., 2021) generates such pairs, and diffusion models expose the mismatch (Huang et al., 2024). MMFakeBench (Liu et al., 2025) extends single-source corpora to mixed-source distortions, and MLLM-based systems add external knowledge and forgery analysis: SNIFFER (Qi et al., 2024), LEMMA (Xuan et al., 2024), FKA-Owl (Liu et al., 2024), HAMMER (Shao et al., 2023).

The current trend wraps an MLLM in a multi-stage verification loop with retrieval, evidence aggregation, multi-round reasoning, and sometimes search-tree planning, as in T^2Agent (Cui et al., 2026). These pipelines report the strongest published numbers at several online LLM calls and retrievals per post, and along the whole ladder accuracy is the only axis of comparison. Per-post cost therefore rises four to five orders of magnitude along this ladder, from single-forward models to agentic pipelines (Figure 6; Table 19). Every system on it optimizes the answer to "real or rumor?", and to our knowledge none can decline when no grounded answer exists. We do not compete on accuracy; we price the alternative, a single-forward detector three to five orders of magnitude cheaper, whose residual errors concentrate exactly where the expensive stage was never needed.

2.2 Selective Prediction: Signals from the Posterior

Selective classification trades coverage for accuracy, from the noise-free formulation of El-Yaniv & Wiener (2010) to deep selective classifiers (Geifman & El-Yaniv, 2017); conformal prediction adds distribution-free guarantees (Angelopoulos & Bates, 2023) on top of algorithmic-learning bounds (Vovk et al., 2005). For LLMs, abstention has been studied via verbalized confidence (Xiong et al., 2024), semantic uncertainty (Kuhn et al., 2023), and self-knowledge elicitation (Yin et al., 2023). Both lines derive the rejection signal from the model's own posterior, which fails for exactly the posts that need refusal (Section 3). Classical selective prediction also terminates on abstention; our gate instead defers rejected posts to a second-level handler (Sections 4.2–4.3).

2.3 Probing and Claim Checking: What Is Decoded

Linear probes recover linguistic structure from frozen representations (Alain & Bengio, 2017), including the classical NLP pipeline (Tenney et al., 2019); Belinkov (2022) surveys their limits. Truthfulness and error states are likewise linearly decodable (Azaria & Mitchell, 2023), including what a model knows versus shows (Orgad et al., 2025) and the geometry behind it (Park et al., 2024). Closest to us, Marks & Tegmark (2024) show a linear *truth direction* and Savcisens & Eliassi-Rad (2025) decode "neither" statements, which the model cannot verify. Both probe the model's knowledge boundary, whereas we probe whether a multimodal post is verifiable at all, something no model's output distribution, however confident, can settle. Because the function we decode concerns the input rather than the model's knowledge, it transfers unchanged across backbones. Check-worthiness detection, from ClaimBuster (Hassan et al., 2017) to the CLEF CheckThat! lab (Nakov et al., 2021), ranks already-extracted claims for a fact-checker's queue, so our decision sits one level earlier and precedes claim extraction.

3 Problem Formulation and the Failure of Confidence

3.1 Task and Detector

A post consists of text with an optional image, and the task is binary classification into *real* (0) or *rumor* (1). We detect with the minimal recipe of Section 1: Qwen3-VL-4B (Bai et al., 2025) frozen, LoRA on the attention projections, and a linear head; our claims concern the *rejection policy*, not the detector. Formally, let $x = (t, I)$ denote a post with text t and optional image I. The detector f maps x to logits $f(x) \in \mathbb{R}^2$ and predicts $\hat{y} = \arg\max_{c\in\{0,1\}} f_c(x)$, whose signed *margin* $m(x) = f_1(x) - f_0(x)$ is the confidence score the baseline gate thresholds ($\hat{y}$=1 iff $m(x) \geq 0$).

We evaluate on four benchmarks: AMG (Guo et al., 2025) (multi-granularity image–text fake news, English), FineFake (Zhou et al., 2026) (English), Weibo (Jin et al., 2017) (Chinese), and MR2 (Hu et al., 2023) (real / rumor / *unverified*; English); AMG and FineFake carry finer sub-labels we merge to a real/rumor binary. Throughout the paper, a post is verifiable when it asserts a proposition checkable against public evidence, and *no-claim* when it asserts none; verifiability denotes this axis, operationalized as an LLM judge's verdict on whether the text contains a verifiable claim (audited in Appendix G). MR2's label-2 (*unverified*) is our *out-of-scope* set but is not identical to no-claim (53.88% of the corpus, 59.5% of the evaluation subset); the handler separates them using the same judge verdict. Splits follow the original papers; thresholds are validation-selected and the test set is touched once.

3.2 The Failure of Confidence

The dominant error mass is *no-claim* content, by which we mean posts asserting no claim to check, such as opinions, jokes, and edited images. On MR2 the unverified (label-2) class is the out-of-scope set and the bulk of what the binary detector misclassifies; these are precisely the posts on which it is confidently wrong. The same reading on the in-domain benchmark is extreme: on AMG test, mean confidence on wrong predictions reaches 0.9372 (Appendix D). Confidence-only rejection barely helps: answered accuracy rises from 0.5596 at full coverage to only 0.5696 at 79% coverage, and 428 of the 506 label-2 posts still leak; a conformal wrapper inherits the blindness (Appendix A).

The asymmetry is quantitative: 53.88% of label-2 posts carry no verifiable claim, roughly $8.47\times$ the real/rumor rate (Appendix F), and in the transfer domains the rate is 10.8–13.7%. Unverifiable content is thus a concentrated failure mode, rare in ordinary content but dominant where the detector must refuse.

Inside the modeled label space, margin is a working difficulty signal: on AMG it detects the detector's own errors at AUROC 0.735 (bootstrap CI in Appendix D), and confidence-based rejection works in-domain, in the sense that discarding the 40% least-confident AMG posts raises answered accuracy from 0.844 to 0.916.

What confidence cannot do is see the boundary of that space. On MR2, posts from the class the detector never models carry higher margin than ordinary posts (7.06 vs. 0.81 in-domain), so the confidence gate refuses 0.0% of them, that is, none, while the verifiability gate refuses 67.1%. The

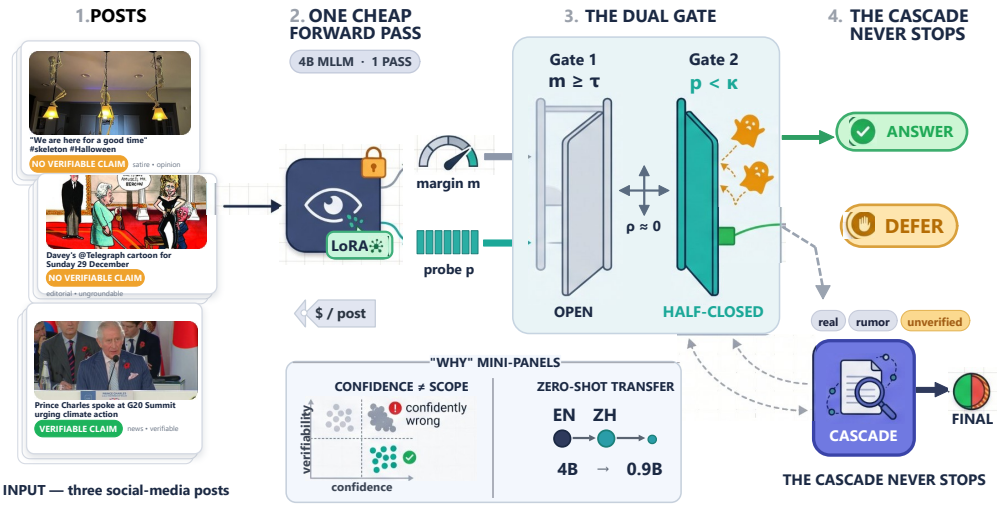


Figure 1: System overview. Three example posts (left) enter a frozen MLLM detector (Qwen3-VL-4B + LoRA head) that emits a margin m and a pooled view in one forward pass; two of them assert no verifiable claim (a satire meme and an editorial cartoon), while one is a news assertion that does. The dual gate answers only if margin $\geq \tau$ *and* probe $p_{\text{nc}} < \kappa$: the two no-claim posts are refused whatever their margin, and the news post passes. Rejection is a deferral, not a termination; refused posts go to a cascade that retrieves evidence and falls back to the detector when none is grounded. The mini-panels show that confidence measures difficulty rather than scope (the confidently-wrong failures sit in the high-confidence, low-verifiability quadrant) and that the probe transfers zero-shot across language (EN→ZH) and scale (4B→0.9B).

signals are near-orthogonal (Spearman $\rho = 0.089$ on AMG, n=874): confidence ranks how hard a post is, while verifiability says whether it is in scope at all. In-domain the detector is not most confident on no-claim content (margin 9.16 vs. 10.99); the confidently-wrong failure is a cross-boundary phenomenon, measurable on MR2.

One may ask whether the probe is repackaged novelty detection, or merely any supervised readout. In-domain the probe separates no-claim from claim at AUROC 0.914 ([0.879, 0.941]) while Mahalanobis is at chance (0.519) and kNN reaches 0.850; over $k \in \{1, \ldots, 20\}$ and PCA dimensions 16–256 neither baseline approaches the probe (0.804–0.866; 0.509–0.705). On AMG the gap widens (kNN 0.55, Mahalanobis 0.65, confidence *inverted* 0.39, against 0.9512). Under an identical converged readout, probes trained on the same representations with detector-error, detector-correct, or real/rumor labels reach only 0.642, 0.375, 0.615 (shuffled 0.356) against 0.912 (Appendix E). The construct is therefore not recoverable from an arbitrary supervised direction.

4 Method

Figure 1 overviews the system: a post is answered only when both the margin gate ($m \geq \tau$) and the verifiability gate ($p_{\text{nc}} < \kappa$) pass; otherwise it is routed to a second-level handler.

4.1 The Verifiability Direction

Verifiability is a property of the *input*, not of any detector, which is precisely why it is the right coordinate for abstention. We obtain no-claim labels by distilling an LLM judge that decides whether the text contains a verifiable factual claim (Appendix P). One caveat bounds everything downstream: the probe decodes the judge's labeling function, so judge biases are inherited; our two-annotator audit (Appendix G) bounds this with Cohen's κ, and the transfer results show the function is stable across

domains, languages, and families. An independent single-sentence rubric reproduces the human majority at only $\kappa = 0.374$, yet the direction remains recoverable under those labels (0.846/0.834 vs. 0.971/0.977, a -0.13 AUROC shift), so the boundary is instrument-sensitive while the direction is not one instrument's artifact. What we claim is thus a human-reproducible convention under a fixed rubric, linearly decodable and deployment-useful, not reducible to lexical statistics (Appendix K).

We pool frozen-detector hidden states into candidate *views* (text-token mean, final token, image-token mean, full-sequence means at several depths) and fit PCA followed by ℓ_2-regularized logistic regression, selecting view and PCA dimension on validation. The readout is deliberately minimal, since if the property is linearly present, a trivial probe should find it. Concretely, for a view $h(x) \in \mathbb{R}^d$ let $z(x) = \mathrm{PCA}_k(h(x))$ be its k-dimensional projection and let σ be the logistic function. The no-claim posterior and its fit are

$$p_{\mathrm{nc}}(x) \;=\; \sigma\big(w^\top z(x) + b\big)\,, \qquad (w^*, b^*) = \arg\min_{w,b} \sum_{i=1}^{n} \ell_{\log}(y_i^{\mathrm{nc}},\, p_{\mathrm{nc}}(x_i)) + \lambda \|w\|_2^2, \tag{1}$$

where $y_i^{\mathrm{nc}} \in \{0, 1\}$ is the distilled no-claim label of post i and the pair (k, λ) is chosen on validation. Nothing beyond this linear readout is used: no probe architecture, retrieval, or target-domain labels.

4.2 Verifiability-Gated Selective Prediction

Given the no-claim posterior $p_{\mathrm{nc}} \in [0, 1]$ and the detector margin m, the two-gate rule answers a post only if $m \geq \tau$ *and* $p_{\mathrm{nc}} < \kappa$ (Figure 1):

$$\mathrm{answer}(x) \;=\; \mathbb{1}[\, m(x) \geq \tau \;\wedge\; p_{\mathrm{nc}}(x) < \kappa \,]\,, \tag{2}$$

where the direction shown is the *forward* setting; the *reverse* setting abstains on low p_{nc} instead, and which one to use is itself a per-domain choice (below). Both thresholds are validation-selected: τ targets 0.95 rumor-bucket precision, κ the same answered-precision through the no-claim posterior (Table 5); test sets are scored once. The two near-orthogonal signals partition posts into four regimes (Table 27, Appendix O). Confidence-only rejection is blind to the no-claim column: the meme in Figure 7(a) is refused by gate 2 no matter how confident the detector is, while verifiable claims can be adjudicated against public evidence.

"Confidence-only" subsumes more than the raw margin. For a two-class head the calibrated posterior $\sigma(m/T)$ is strictly increasing in m, so quantile thresholds on a calibrated score induce the *same* rejection set as on m: temperature scaling and conformal wrappers are bit-identical in routing (Appendix A), and the label-uncertainty family (MC-dropout, ensembles) ranks the same posts by construction, since spread over labels cannot see a failure that is a property of the input. The non-degenerate input-side variant, paraphrase consistency, inverts: out-of-scope posts are the most rewrite-stable (0.99 vs. 0.96), and its strictest gate still leaks 488/506 label-2 posts at nine-fold cost; training-free input cues fall far below the probe too (0.742 vs. 0.9512; Appendices O, N).

The gate direction (no-claim $\Rightarrow$ abstain, or reinforce the rumor call) is a deployment choice we set automatically from a sign criterion: the no-claim rates of validation FPs vs. TPs. On AMG no-claim is over-represented among FPs (35.7% vs. 11.1%), so no-claim $\Rightarrow$ refuse; on FineFake and Weibo the pattern reverses (2.8% vs. 21.8%; 0.0% vs. 6.4%), so no-claim $\Rightarrow$ rumor. The gate thus reads off each benchmark's own convention for unverifiable content (three correct predictions, no failures; an empirical heuristic, not a theorem). The probe transfers zero-shot, whereas the direction, read off the same gold-labeled validation bucket that calibrating τ already requires, is a per-domain operating choice with no single decision semantics across benchmarks (no extra annotation, but not zero-shot either; forcing a source-fixed direction collapses FineFake, Appendix N). It also explains why the detector cannot supply the signal itself: fine-tuning encodes the rumor criteria its data treats as universal, whereas unverifiable posts are boundary samples on which benchmarks disagree, and the validation split reads off that disagreement.

4.3 Post-Abstention Cascade

We route rejected posts to a second-level handler that always returns a final judgment, deciding where it can and falling back to the detector where it abstains, so abstention triggers a second level rather than terminating, giving a cascade (prompts in Appendix P). On the binary domains the handler is a retrieval-capable agent issuing a 0/1 judgment after search, abstaining to the detector when

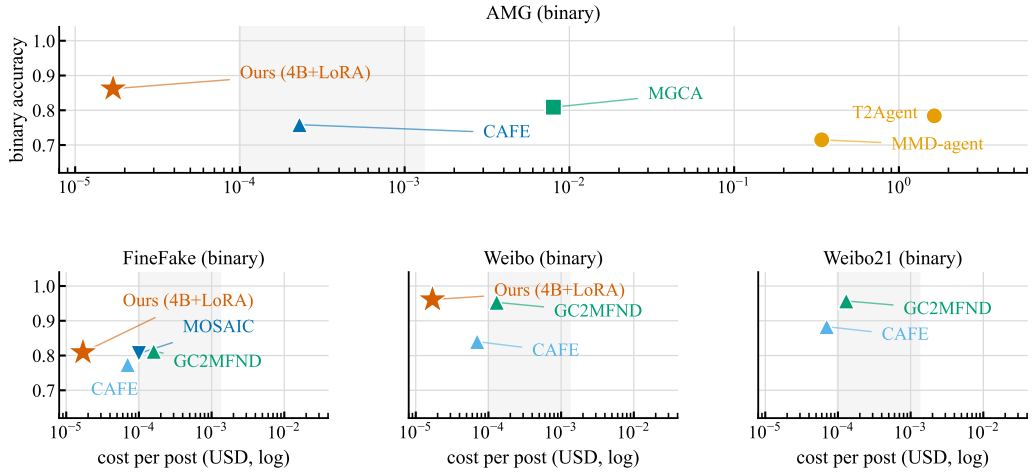


Figure 2: Inference cost per post (USD, log-x) vs *binary* accuracy, one panel per dataset; the red star is our detector (0.862/ 0.809/ 0.961) at 1.7×10^{-5} USD per post, with no online LLM calls or retrieval (Appendix Q). Accuracy is comparable only *within* a panel and every non-star point is as reported by its source paper, so the cost axis mixes accounting protocols and no comparison here is like-for-like. The AMG panel is a protocol-mismatched anchor and Weibo21 is not one of our benchmarks, so that panel only positions the published systems. The grey band marks the $\sim 10^{-4}$–10^{-3} single-local-forward range.

evidence is insufficient; on MR2 a rule cascade answers claim-free types as unverified (label 2), forces a 0/1 choice on news-category types, and uses a three-way prompt for the rest.

5 Experiments

5.1 Experimental Setup

Detection, labeling, splits, and the train-fit/validation-select/test-once protocol are as in Sections 3–4.2; the probe's view and PCA dimension come from the grid in Appendix H, and both gate thresholds are validation-selected targeting 0.95 rumor-bucket precision. Metrics are AUROC for the probe and accuracy/precision/recall/F1 for detection and routing. Prompts, cost protocol, and statistics are in Appendices P, Q, M.

5.2 Main Results

Under the original splits the minimal recipe reaches 0.862 (AMG, three-seed mean), 0.809 (FineFake), and 0.961 (Weibo) test accuracy (Table 18), competitive with published specialized detectors (protocol caveats in Appendix L). This is not a SOTA claim but a recalibration: when accuracy is this cheap, the residual failure is not an accuracy problem, because it concentrates on no-claim content where more accuracy cannot help.

Figure 2 plots per-post cost against binary accuracy (comparable only within a panel). On AMG our detector (0.862) sits above the agentic pipelines (T2Agent (Cui et al., 2026) 78.40, MMD-agent 71.50) and MGCA (80.89) at 10^2–$10^5\times$ lower cost; on FineFake we trail the published systems by 2–3 points, and on Weibo only RoViT (98.93) is above us. Every comparison is on the binary real-vs-fake dimension each number reports, and we claim no state of the art, since the contribution is the rejection layer none of these systems provide. Table 1 reports the 0.95-precision operating point. On AMG the forward gate buys 8 extra intercepted FPs (86 of 126) and +0.6 answered accuracy for 5.1 coverage points; an oracle gate intercepts 91/126, so of the interceptions the oracle adds over confidence-only the probe captures 62% at the 0.95 target (8 of 13) and 73% at 0.90 (19 of 26). Rejects are error-concentrated (forced to answer they would err at 54.8%, 8× the answered rate, Fisher $p < 10^{-30}$) but costly: 45.2% of the 157 rejects would have been answered

Table 1: Selective prediction at the 0.95-precision operating point (validation-selected thresholds; test scored once). Gate direction follows the sign criterion: *forward* = abstain when p_{nc} is high (no-claim ⇒ refuse), *reverse* = abstain when it is low (no-claim ⇒ rumor); only rumor-side predictions ($\hat{y}$=1) can be abstained, so coverage is bounded below by the real-predicted share. "val FP/TP no-claim" is the sign-criterion statistic shared by both rows; The capture ratio ("62–73% of the increment") is (gate − conf)/(oracle − conf) in intercepted FPs at the two precision targets: $(86-78)/(91-78) = 8/13$ at 0.95 and $19/26$ at 0.90. McNemar at the matched point: AMG $p = .057$, FineFake $p = .0014$, Weibo $p = 3.1 \times 10^{-5}$; the nested pinned-τ variant is significant in all three.

Domain	gate direction	val FP/TP no-claim	coverage	acc (answered)	FPs intercepted
AMG	confidence only	35.7% vs. 11.1%	87.1%	0.924	78/126
	+ verifiability gate (fwd)		82.0%	**0.930**	**86/126**
FineFake	confidence only	2.8% vs. 21.8%	74.3%	0.892	186/211
	+ verifiability gate (rev)		52.2%	0.866	**204/211**
Weibo	confidence only	0.0% vs. 6.4%	99.3%	0.967	6/23
	+ verifiability gate (rev)		60.4%	**0.974**	**22/23**

correctly. On FineFake and Weibo the reversed gate makes the direction a rumor-side signal (precision 0.956→0.965, 0.963→0.989) at heavy coverage cost (74.3%→52.2%; 99.3%→60.4%), with 74.7%/94.1% of rejects answerable. On MR2 the rule gains $+2.8$ points at matched 95% coverage and reaches 0.6942 at 75%, beyond the confidence-only frontier (Appendix B). Because the rule *defers* rather than classifies, the shared event is the rejection itself: FineFake and Weibo are significant ($p = .0014$, $p = 3.1 \times 10^{-5}$), AMG is not ($p = .057$; only the nested pinned-τ variant is, Appendix M).

On AMG, FineFake, and Weibo the fully-answered pipeline exceeds the detector's direct accuracy at the same seed (84.4%→86.04%, 80.9%→83.20%, 96.1%→96.44%), which are small *coverage* gains rather than accuracy gains. An alternative operating point is reject-only, abstaining wherever the gate abstains and never invoking the handler. On the binary domains that is a high-precision, low-coverage point (AMG 93.0% at 82.0%) whose accuracy exceeds the full cascade's, buying coverage back at 57× the per-post cost; on MR2 the sign flips (reject-only 0.750 vs. 0.812). A trivial-fill bound (Appendix N) shows the cascade beats every constant or random fill on FineFake and Weibo but sits 0.1 points below the strongest on AMG, so reject-only stays on the table and the gate's job is deciding which posts incur the second level's cost.

MR2's label-2 (*unverified*) is our only genuine three-class scenario, and the binary-trained detector has no boundary for a third class: the confidence gate refuses 0.0% of it, the verifiability gate 67.1%, while ordinary 0/1 posts are refused at only 7.7% (Appendix B). Direct three-way accuracy is 0.560; a conformal gate alone recovers to 0.729, and adding the verifiability criterion reaches 0.812 (Table 2), a figure carried by the no-claim half of label-2, which we decompose next.

The 0.812 figure is interpretable only once the label-2 set is split by its own judge labels (Appendix C): of the 506 posts, 301 (59.5%) are genuinely no-claim and 205 carry an unconfirmed claim. The confidence gate refuses 0.0% of *each*; the verifiability gate refuses 72.4% of the no-claim half but only 29.8% of the claim-bearing half, scoring 1.00 vs. 0.298. The headline is thus carried by the no-claim half, where refusal is correct; the claim-bearing half shows no over-refusal, only the limit of a verifiability gate on merely unresolved posts.

5.3 Ablation Studies

All ablations reuse the same detector, splits, and validation-selected thresholds, so a gap is attributable to the component alone. Table 2 reads as a gate-component ablation: a conformal gate alone recovers 0.729, adding the verifiability gate reaches 0.812, removing the handler's evidence drops to 0.794, and replacing the LLM-distilled gate with a self-contained probe drops to 0.758(bootstrap CIs in Appendix M). The view and (k, λ) surface is swept in Appendix H (selected cell ranks 3 rd of 96; text-side pooling dominates image-side, Figure 5), the shallow-text control rules out a lexical proxy (Appendix K), and the label-swap control rules out an arbitrary supervised direction (Table 9).

Table 2: Rejection behavior on MR2 (merged test + out-of-scope set; all gates use validation-selected thresholds). MR2's label-2 is a class the detector never models (it is trained on 0/1 only). Accuracy is three-way over *all* posts: an abstention scores as label 2, and counts as correct when the gold class is 2, unlike Table 5 whose accuracy is computed over the answered subset only.

System variant	accuracy
Detector direct 3-class (no rejection)	0.560
+ conformal rejection gate (no handler)	0.729
+ verifiability gate (**full system**)	**0.812**
Ablations:	
no evidence available to handler	0.794
self-contained gating (probe replaces LLM-distilled gate)	0.758

5.4 Analysis: Properties of the Verifiability Direction

On AMG (seed 42) the probe reaches val AUROC 0.9344 and test AUROC 0.9512; across three detector seeds (view fixed to the text-token mean), 0.9473 $\pm$ 0.0054 (Appendix J). The selected cell is not lucky: over the full 96-configuration selection surface the validation-selected cell ranks 3 rd, and no cell exceeds it by more than noise (Appendix H). Text-side pooling is the right carrier, as the text-token mean view reaches 0.95 while the image-token view never exceeds 0.79, so the construct, defined on the text, is carried by the claim text rather than the pixels (a definitional consequence of the text-side rubric, not an image-side finding; Appendix H).

The AMG-trained head, applied unchanged to FineFake and Weibo representations, reaches 0.9827 $\pm$ 0.0007 and 0.9583 $\pm$ 0.0033 over three seeds (Table 15). Weibo is Chinese, so the transfer is cross-lingual, without translation or target labels; a same-domain head fit on small target validation sets shows the direction is not source-specific (Appendix I).

A linear readout could in principle proxy lexical surface statistics. Under an identical protocol (TF-IDF n-grams, best of 108 validation-selected configurations; Appendix K) the best shallow model reaches 0.860 in-domain against the probe's 0.9512, and collapses to 0.530 on cross-lingual Weibo while the probe holds 0.9583. Refit with target labels it still stays below the zero-shot probe (0.975/0.925 vs. 0.9827/0.9583), and concatenating TF-IDF onto the representation adds only $+0.002$.

Repeating the protocol on three further MLLMs (Qwen2.5-VL-3B, SmolVLM2-2.2B, LLaVA-OneVision-0.5B), each with its own detector, labels, and independently validation-selected view and PCA dimension, per-backbone test AUROC spans 0.8886–0.9512 (Table 16): family breadth, not scale, is what matters.

The direction is present at every layer of the 37-layer Qwen3-VL-4B detector (Figure 5): from the first layer (AUROC 0.80) it rises monotonically to $\approx$0.965 at layers 23–25. Repeating on the frozen backbone gives an indistinguishable curve (mean per-layer $|\Delta\mathrm{AUROC}| = 0.007$, comparable to the seed-to-seed probe spread ± 0.0054), so, at this resolution, the direction is a property of the pretrained space that fine-tuning neither creates nor reshapes, consistent with its zero-shot transfer.

5.5 Analysis: A Controlled Negative Result

We test the causal story as *subspace-orthogonal regularization*: an ℓ_2 penalty on the projection of the classifier readout onto a fixed k=256-dimensional orthonormal subspace V, comparing V = the verifiability direction against a random subspace of the same dimensionality across three AMG seeds, the critical control being that a benefit in both arms cannot be attributed to the direction's semantics (FineFake and Weibo repeat the verifiability arm only).

Across the three AMG seeds (Table 23) blocking the verifiability direction is indistinguishable from baseline (0.8608 ± 0.0270, high seed variance), while blocking a *random* subspace is stably better (0.8795 ± 0.0007): the benefit is a *regularization* effect under distribution shift, and the confidently-wrong errors are *not* caused by reliance on this direction.

The two-domain extension (FineFake/Weibo) is consistent (Appendix M); the direction still carries semantics, as on MR2 blocking it collapses label-2 recall from 0.88 to near zero. It is worth *reading*, not *blocking*. Two claims must be kept apart: the direction is *not* the cause of the confidently-wrong errors, but it *is* the basis of the gate's rejection decision, which reads it directly. Why can a non-causal direction still ground rejection? Rejection sits at a different decision level from prediction, being a normative rule about which posts should be answered at all, and this is what makes a non-causal marker actionable. On the binary domains the label boundary does not turn on verifiability, so reliance on the direction is incidental; on MR2 the third class is precisely the unverifiable content, so blocking removes a task-relevant feature. A no-claim post goes to a human or to the platform's policy layer, since no model improvement can ground a verdict on content that asserts nothing verifiable, whereas posts carrying an unconfirmed but verifiable claim are answered, at low accuracy (Appendix C); a deployment with asymmetric error costs can overlay a confidence or source-credibility overrule on that half without touching the gate.

6 LIMITATIONS

Our first limitation concerns label provenance. No-claim labels are LLM-distilled, so judge biases are inherited; the two-annotator audit (Appendix G, n=176) bounds this ($\kappa = 0.850$ inter-annotator, $\kappa = 0.804$ distilled-vs-majority) but is weaker on the no-claim class the gate depends on (0.812), and an independent-rubric re-audit drops agreement to $\kappa = 0.374$, so we claim a convention under a fixed rubric rather than a natural kind. Second, the gate is not free (45.2%/74.7%/94.1% of AMG/FineFake/Weibo rejects would have been answered correctly) and reject-only is a legitimate alternative (Appendix N); the cross-system table mixes protocols, the handler is evaluated on MR2 only, and label-2 is not a pure no-claim set (Appendix C). The AMG matched-point test is not significant ($p = .057$), the sign criterion is an empirical heuristic, transfer is EN→ZH only, and the random-subspace control runs on AMG only. Whether more in-domain training data would narrow the confidently-wrong gap is untested.

7 CONCLUSION

Confidence tells a detector how sure it is, not whether there is anything to be sure about. A verifiability signal is linearly decodable from multimodal LLM representations, transfers across datasets, languages, and model families, and a gate built on it restores selective prediction where confidence fails. The lesson for deployment is to check that the rejection signal can see the failure mode one actually has. Future work should address multi-label and multimodal claims, where a post may carry both verifiable and unverifiable assertions, and adapt the criterion as benchmark conventions shift.

AI USE STATEMENT

In this work, we used generative AI tools for one task that requires disclosure under the ICLR 2027 AI Policy for Authors: producing the synthetic no-claim labels that serve as supervision and as the analysis target throughout the paper, where a generative model acts as a judge under a fixed rubric (Section 4.1, Appendix P). We did not use generative AI tools to implement the methods, to formulate the mathematical claims, to construct, check, or supply critical ingredients for any proof, to propose or refine the hypotheses, to develop the theoretical model or the conceptual framework, to design the methodology or the experiments, or to interpret the reported results. The remaining tasks requiring disclosure (assisting with translation, cleaning or reformatting datasets, and supporting qualitative or thematic analysis) were not part of this work, and no other required-disclosure subtask applies to this submission. Additionally, we used generative AI tools for tasks where disclosure is recommended: language editing of the manuscript, and sourcing and searching for related literature. We have reviewed all AI-assisted work: all text and code were checked by the authors, every reported number traces to a logged run, and the judge-derived labels used as supervision were validated against a two-annotator human audit (Appendix G). We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

All experiments use fixed data splits, seeds (42/43/44), and a train-fit/validation-select/test-once protocol. Appendix P gives the exact judge and agent prompts and rubrics, together with the judge/handler model endpoint and the retrieval backend; Appendices H and K give the hyperparameter grids; Appendix M gives the statistical procedures and per-seed metrics (three-seed detector and transfer values; the per-seed numbers ship with the release); Appendix N gives the direction-control, heuristic-baseline, and trivial-fill recomputations; Appendix G gives the human audit and the independent-rubric labels. Code and distilled labels will be released.

Ethics statement

This work targets misinformation triage; abstention is a safety mechanism, and false refusal has a measured cost we report (coverage). It also carries a governance risk we flag: refusing no-claim content (jokes, satire, personal posts) can shade into over-censorship, which is why we report the coverage cost and keep the no-claim verdict distinct from the handling decision rather than hiding it behind a single accuracy number. LLM-distilled labels are human-audited. We do not release identity-bearing data; all datasets are public research benchmarks.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Workshop Track Proceedings*, 2017.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):7–35, 2022.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pp. 2897–2905, 2022.

Xing Cui, Yueying Zou, Zekun Li, Peipei Li, Xinyuan Xu, Xuannan Liu, and Huaibo Huang. T^2Agent: A tool-augmented multimodal misinformation detection agent with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 175–183, 2026.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 228–236, 2025.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1803–1812, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2901–2912, 2023.

Mingzhen Huang, Shan Jia, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Exposing text-image inconsistency using diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, 2017.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the 2019 World Wide Web Conference*, pp. 2915–2921, 2019.

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

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. FKA-Owl: Advancing multimodal fake news detection through knowledge-augmented LVLMs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10154–10163, 2024.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. MMFakeBench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, 2021.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mücahit Kutlu, and Yavuz Selim Kartal. Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the Cross-Language Evaluation Forum (CLEF 2021)*, pp. 264–291. Springer, 2021.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kiho Park, Yo Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *First Conference on Language Modeling*, 2024.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024.

Germans Savcisens and Tina Eliassi-Rad. The trilemma of truth in large language models. *arXiv preprint arXiv:2506.23921*, 2025.

Table 3: Temperature calibration: statistical calibration improves NLL/ECE but leaves routing decisions unchanged in all three domains.

Domain	T	NLL (test, pre→post)	ECE (pre→post)	routing decisions
AMG	2.93	$1.018 \rightarrow 0.443$	$0.145 \rightarrow 0.151$	unchanged
FineFake	1.60	$0.415 \rightarrow 0.396$	$0.078 \rightarrow 0.060$	unchanged
Weibo	2.02	$0.123 \rightarrow 0.102$	$0.029 \rightarrow 0.022$	unchanged

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913. IEEE, 2023.

Shivam Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 75–83. IEEE, 2019.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 849–857, 2018.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2560–2569, 2021.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. LEMMA: Towards LVLM-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

Xiaofan Zheng, Minnan Luo, and Xinghao Wang. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7862–7869, 2025.

Ziyi Zhou, Xiaoming Zhang, Litian Zhang, Jiacheng Liu, Erik Cambria, and Chaozhuo Li. FineFake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection. *Information Fusion*, 132:104253, 2026. doi: 10.1016/j.inffus.2026.104253.

A Calibration Does Not Fix the Failure

We fit a temperature T on validation (all three domains are over-confident: $T = 2.93/1.60/2.02$ for AMG/FineFake/Weibo) and re-run the entire routing evaluation with calibrated MSP in place of the raw margin. Calibration improves NLL and ECE as expected, but *every routing decision is unchanged*: coverage, rumor-bucket precision, and intercepted FPs agree bit-for-bit with the uncalibrated run in all three domains.

Table 4: Rejection-source statistics on MR2. "conf." is the confidence (margin) gate, "verif." the verifiability (no-claim) gate. OOS includes label-2 posts from all splits; the train split's label-2 posts never enter training and are therefore out-of-scope. The larger OOS sample (n=6314) supplements the evaluation set used in Table 5 (n=506, val+test label-2).

Split	total	rejected	rate	rejection source
val	701	135	19.3%	conf. 9.8% + verif. 10.0%
test (0/1)	802	62	7.7%	verif. 7.6% (conf. $\approx$ 0)
OOS (label-2)	6314	4237	67.1%	all from verif.

Table 5: Risk–coverage on the MR2 merged set (1308 posts; accuracy is computed over the *answered* subset at each coverage level, as abstained posts are excluded rather than scored as errors, so accuracy can and does exceed coverage in the low-coverage rows). The verifiability gate is the AMG transfer head, applied zero-shot. "Label-2 answered" counts unverified (label-2) posts that leak into the answered set. The confidence-only curve is measurable only down to $\approx 79\%$ coverage, so rows below that report the two-gate accuracy and its leakage alone.

Coverage	confidence-only answered acc	+ verifiability gate answered acc	label-2 answered (conf $\rightarrow$ dual)
100%	0.5596	–	506 $\rightarrow$ –
95%	0.5611	0.5886	492 $\rightarrow$ 441
79%	0.5696	–	428 $\rightarrow$ –
75%	–	0.6942	– $\rightarrow$ 230
53%	–	0.7682	– $\rightarrow$ 94

The reason is structural: in binary classification, softmax of the scaled logits is a monotone function of the margin, so quantile thresholds on either score induce identical rejections. The practical consequence is sharper: AMG's threshold-transfer degradation (rumor-bucket precision -0.15 from val to test) persists *unchanged* after calibration, while FineFake/Weibo are stable (± 0.01); the degradation stems from semantic distribution shift in the test rumor bucket, not from a confidence scale drift. Statistical calibration cannot repair it; a semantically grounded second signal (the verifiability direction) is required.

B Rejection-Source Statistics

We quantify the gate mechanism at the sample level on MR2, using the detector's own margin (confidence gate) and no-claim posterior (verifiability gate). The detector is trained on MR2's real/rumor (0/1) labels only; the out-of-scope sample below includes label-2 posts across all splits, none of which enter any training step (they are used purely as out-of-scope test items). We report how many posts each split rejects and from which gate the rejection originates.

The pattern is the mechanism in miniature: on out-of-scope content the confidence gate rejects almost nothing (its scores are confidently high), while the verifiability gate rejects 67.1% of the same posts; on real/rumor test posts only 7.7% are rejected, so the gate does not harm normal content. Because these OOS items include label-2 posts the system never trained on, whose no-claim subset (Section 3) is genuinely unverifiable, the rejection behavior transfers to unseen refusal-target content without target-domain training.

Table 5 complements these sample-level statistics with the MR2 risk–coverage scan used in Section 4.2.

C Decomposing the Unverified (Label-2) Set

MR2's label-2 class mixes two things: content with no verifiable claim, and content whose claim the annotators could not confirm. Treating the whole class as out-of-scope risks rewarding over-refusal if the gate simply rejects everything labelled 2. We therefore split the n=506 evaluation label-2 posts (val+test, the set behind Table 5) by the judge's own `has_verifiable_claim` field: 301

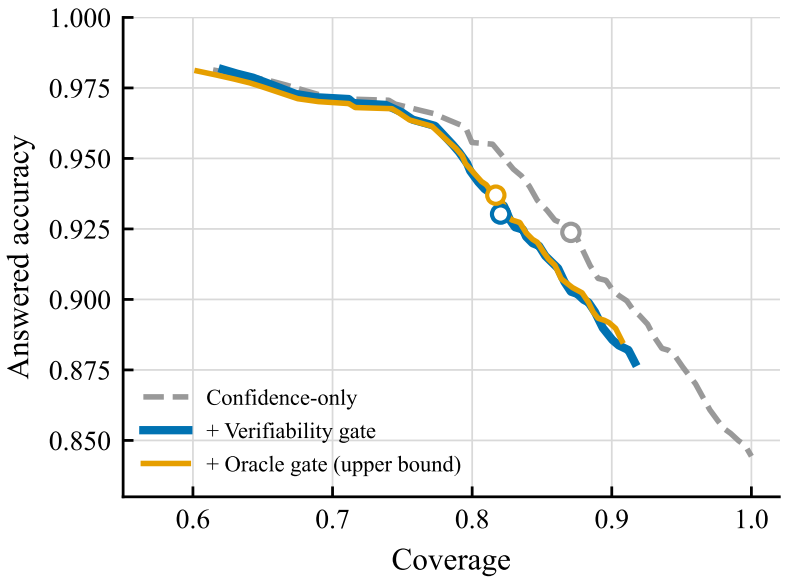


Figure 3: Risk–coverage curves on AMG: confidence-only rejection vs. the two-gate rule, with an *oracle* no-claim gate as upper bound. The oracle gate is not deployable, because it uses the LLM judge's `has_verifiable_claim` ground-truth label as the reject signal, i.e. a perfect verifiability detector, and it bounds how much any learned gate could intercept. The two-gate curve dominates confidence-only in the coverage region the two curves share; at the 0.95 operating point it holds 0.930 answered accuracy at 82.0% coverage vs. 0.924 at 87.1% for confidence-only.

Table 6: The label-2 evaluation set (n=506) split by the judge's own no-claim label. The confidence gate refuses neither half; the verifiability gate refuses the no-claim half far more than the claim-bearing half; three-way accuracy (abstention scored as label 2) is *not* uniformly high, and is carried entirely by the no-claim half; on that half 1.00 is partly by construction, since refused posts score as label 2 and the rule cascade answers the surviving no-claim posts as label 2.

Subset	n	conf. gate refuses	verif. gate refuses	mean margin	3-way acc
no-claim	301	0.0%	72.4%	10.09	1.00
claim-bearing (unverified)	205	0.0%	29.8%	9.53	0.298

posts (59.5%) are genuinely no-claim and 205 carry a claim left unverified. Table 6 reports the two gates' refusal behaviour and the full system's three-way accuracy on each half.

The picture is the opposite of over-refusal, and the two halves reach their figures by different routes. On the no-claim half every post ends up labelled 2: the 72.4% refused by the gate score as abstentions, and the remaining 83 of 301 are answered by the rule cascade, which maps claim-free content to *unverified* by construction; hence 1.00, which is therefore partly an artifact of the scoring convention (the handler rule that assigns label 2 to no-claim posts mirrors the annotation convention that created the gold labels). On the claim-bearing half the rule does not fire: the verifiability gate refuses only 29.8%, correctly recognising that these posts *do* contain claims, and the system answers the rest, so the figure is exactly the abstention share (29.8%, matching 0.298) and the remaining $\approx$70% are answered with a $0/1$ verdict and are all wrong. The 0.298 is thus not a mix of over- and under-refusal. In other words, the 0.812 headline is not a product of blanket refusal: it is produced by the no-claim half, and the claim-bearing half is where the system (by design) declines to abstain. The residual 0.298 is a genuine limit, but a different one from over-refusal: it reflects that a verifiability gate has no opinion about whether an unconfirmed claim will eventually be resolved, and MR2's scoring penalises that. A deployment can of course add a confidence or source-credibility overrule on that half; we keep the gate's semantics narrow.

Table 7: Selective-prediction diagnostics on AMG (seed 42; n_{val}=496, n_{test}=874). Lower AURC/e-AURC is better; higher error-detection AUROC is better. All scores are derived from the detector's own posterior; none sees the verifiability signal.

Score	split	AURC	e-AURC	err-detection AUROC
MSP	val	0.0327	0.0301	0.801
	test	0.0992	0.0864	0.731
margin	val	0.0220	0.0194	0.805
	test	0.0702	0.0573	0.735

Because the evaluation subset (59.5% no-claim) is more no-claim-heavy than the label-2 corpus as a whole (53.9%, Appendix F), the same decomposition on the full pool moves the numbers by a few points in the direction of more claim-bearing content; the qualitative reading is unchanged.

D Selective-Prediction Diagnostics

We quantify the "confidently wrong" failure with a diagnostic on the AMG selective-prediction analysis (Appendix D is the source of the 0.9372 figure cited in Section 3). For each test post we record the detector's confidence on its prediction; *mean confidence on wrong predictions* measures how confidently the model errs, and on AMG test it reaches 0.9372. We also report the area under the risk–coverage curve (AURC), its excess over the oracle curve (e-AURC), and the error-detection AUROC of the confidence score. Table 7 reports the diagnostics for the two scores used by the gate (MSP and margin); logit-based variants (max-logit, energy) reach at most 0.746 error-detection AUROC on test. These diagnostics substantiate a claim we now state in its precise form. In-domain, confidence is a *working* difficulty signal: on AMG the margin detects the detector's own errors at AUROC 0.735 (95% CI $[0.688, 0.778]$), so 0.735 is far from chance and we do not describe it as such. What fails is not ranking but *scope*: the same score cannot separate in-space from out-of-space content, and on MR2 it is inverted (Section 3). The 0.9372 figure is therefore a reading on the posts the detector gets wrong, and its interpretation is domain-conditional: it is the cross-boundary case, not a statement about no-claim content in-domain.

E Verifiability vs. Density Baselines: Full Details

Section 3 asks whether the verifiability probe is merely repackaged novelty detection. We test this in the setting where it should matter most, the same news domain, where the only difference between posts is whether a verifiable claim is present. On MR2's in-domain validation posts (which mix real and rumor news but share a domain), we fit each density baseline on the *claim-bearing* posts only and ask whether the no-claim posts are detected as outliers (n=669, of which 65 are no-claim). We control the ill-conditioning that otherwise corrupts high-dimensional Mahalanobis (with 2560 features and only a few hundred fit samples, a raw covariance inverse manufactures spurious separation) by reducing all methods to a common PCA-64 basis. Table 8 shows the result, and we repeat it on the merged val+test set (n=1471, 96 no-claim) to confirm robustness. The MR2 label-2 stress test in Section 4.3 superficially flatters density methods (Mahalanobis reaches 1.0 there), but that separation is an artifact: label-2 is a *different semantic domain* (memes, ads, personal posts) that any density method separates trivially by domain, and the raw 2560-dimensional Mahalanobis that produces the 1.0 is the same ill-conditioned estimate that collapses to chance in-domain.

Sampling uncertainty and hyperparameter sensitivity. Because this comparison is close on paper, we report the uncertainty around it. A paired bootstrap over the MR2 validation posts ($10{,}000$ resamples) puts the probe at AUROC 0.914 with 95% CI [0.879, 0.941]. Sweeping the density baselines' own hyperparameters ($k \in \{1, 3, 5, 10, 20\}$ for kNN and PCA dimension $\in \{16, 32, 64, 128, 256\}$ for both), kNN spans 0.804–0.866 (its best case is at the lowest dimension) and Mahalanobis spans 0.509–0.705, with no setting coming near the probe; the gap is therefore not an artifact of the PCA-64 choice we report in Table 8. We also note the comparison is not symmetric by construction: the probe is supervised on AMG, whereas the density baselines are unsupervised and fit on the MR2 claim-bearing posts. The controls below address exactly this.

Table 8: In-domain detection of no-claim posts, comparing the verifiability probe against label-free density baselines on the same frozen representations (all methods reduced to a common PCA-64 basis to avoid high-dimensional ill-conditioning). Density methods are fit on claim-bearing posts only. MR2 columns: probe applied zero-shot AMG→MR2, density fit on the evaluated split's claim posts. AMG column, the probe's home domain, where density is fit on claim-bearing *validation* posts (basis fit on validation), evaluated on the judge-labeled test subset (n=871; the probe's full-test value there is 0.9512). †Confidence was computed on the val split only (at chance there). The MR2 label-2 comparison is omitted because it flatters density methods trivially: label-2 is a different semantic domain.

Method	MR2 (val, n=669)	MR2 (val+test, n=1471)	AMG (test, n=871)
Verifiability probe	**0.911**	**0.914**	**0.953**
kNN-OOD (k=5)	0.826	0.850	0.55
Mahalanobis	0.509	0.519	0.65
Detector confidence	0.514	n/a†	0.39

The source domain: AMG, and gate-level behavior. The AMG column of Table 8 extends the comparison to the probe's home domain, where the density baselines degrade further: kNN spans 0.55–0.58 over $k \in \{1, 3, 5, 10, 20\}$ and Mahalanobis reaches 0.65, against the probe's 0.953, and detector confidence is *inverted* (0.39), consistent with Section 3 (in-domain, confidence ranks difficulty, not scope). Whatever separates no-claim from claim posts in AMG, it is not distance from the claim-post distribution. We also ran the comparison at the *gate* level: for each method we select the refuse threshold on validation under the same 0.95 rumor-bucket-precision target used for κ (Appendix N gives the routing rule) and score test once. The kNN gate intercepts more FPs than the probe gate (118 of 126 vs. 86) but only by refusing 65.8% coverage (16 points below the probe gate) and lands just below the precision target (0.948); the Mahalanobis gate's threshold does not transfer (test rumor-bucket precision 0.854). Density scores track distributional unusualness, and a gate built on them buys interception the same way a blanket coverage cut does; the probe's interceptions concentrate on the no-claim boundary instead.

Is the probe just *any* supervised signal? A label-swap control. The fair question is not whether a supervised probe beats an unsupervised density estimator, but whether the no-claim direction is distinguishable from *other* supervised directions fit on the same representations. We therefore train five probes under the identical protocol (same AMG train representations, same `txt_mean` view, same PCA + ℓ_2 grid selected on the AMG validation split), differing only in the training label: the no-claim label, the detector's own error indicator, its correctness indicator, the real/rumor binary label, and a randomly permuted label. To make the five arms exactly comparable, all use a single closed-form ridge readout (no iterative fitting, hence no optimisation randomness between arms). Under this solver the verifiability probe reproduces every headline AUROC (AMG test 0.955, MR2 in-domain 0.912, FineFake 0.980, Weibo 0.946), which is also an independent robustness check on the reported numbers. Table 9 gives the result: the shuffled-label probe is at chance (own-val 0.505), detector-error and detector-correct probes reach only 0.642 and 0.375, the real/rumor probe 0.615, and the no-claim probe 0.912. Verifiability is therefore recovered only by the probe trained for it: it is not the generic "detector error" signal in disguise, the damaging result that we do not find, and the real/rumor axis shares only limited variance, consistent with the two gates composing rather than being orthogonal.

The control also explains why an auxiliary error-detection head cannot substitute for the probe: the detector's own error indicator is only weakly learnable from these representations (own-val 0.660) and transfers poorly to the no-claim target, whereas the no-claim label is both learnable (0.952) and transferable.

F MR2 No-Claim Distribution

Figure 4 visualizes the concentration of unverifiable content: the no-claim rate jumps from the real/rumor baseline to 53.88% in the unverified (label-2) out-of-scope class, while in the transfer domains it stays at 10.8–13.7% of ordinary content.

Table 9: Label-swap control on MR2 in-domain no-claim separation (val; n=669, 65 no-claim). All probes share one representation, view, and readout (closed-form ridge on PCA features); only the training label differs, and each probe's own (k, λ) is selected on AMG validation using *its own* label. "own-val" is that probe's validation AUROC on its own label; the last column is the MR2 no-claim AUROC, the quantity of interest.

Training label	own-val	MR2 no-claim AUROC	label rate
no-claim (the probe)	0.952	**0.912**	0.09
detector error	0.660	0.642	0.02
detector correct	0.380	0.375	0.98
real/rumor binary	0.628	0.615	0.36
random (shuffled)	0.505	0.356	0.09

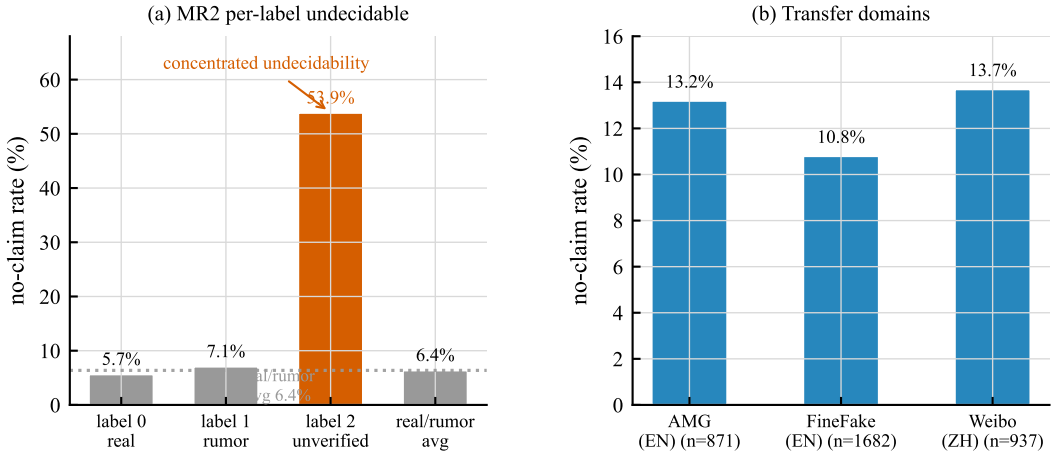


Figure 4: Unverifiable content is a *concentrated* failure mode, not a uniform rate. (a) On MR2, the no-claim rate jumps from 5.7%/7.1% in the real/rumor labels to 53.9% in the unverified (label 2) class that serves as the out-of-scope set; the detector is asked to refuse exactly where unverifiable content dominates, and the dotted line marks the 6.4% real/rumor average. (b) In the transfer domains used for zero-shot evaluation, no-claim posts make up 10.8–13.7% of ordinary content.

Section 3 reports that unverified (label-2) posts in MR2 are far more likely to carry no verifiable claim than real/rumor posts (the 53.88% rate and 8.47 ratio cited there). This appendix reports the per-class no-claim rates that underlie that asymmetry (Table 10). The asymmetry is the quantitative basis for treating unverifiable content as a concentrated failure mode.

G Human Audit of Distilled Labels

Protocol. We draw a stratified sample of n=176 posts, stratified by domain $\times$ distilled label $\times$ detector correctness (the cells where judge error matters most: no-claim labels on detector FPs and TPs in each domain). All domains, including Weibo, are distilled with the same English-language structural-analysis prompt (Appendix P), applied uniformly regardless of the post's language; Weibo items are audited by bilingual annotators, all other domains by English-capable annotators. two annotators independently judge *has_verifiable_claim* against the rubric of Section 4.1 (truth-apt proposition, checkable against public evidence), blind to the distilled label and to each other. We report: (i) inter-annotator agreement (Cohen's κ); (ii) agreement between the distilled labels and the human majority vote; and (iii) the per-domain breakdown of both.

Results. Two independent annotators agree on 0.926 of the 176 posts (Cohen's $\kappa = 0.850$), an *almost-perfect* level by Landis–Koch that confirms the rubric names a construct humans can label reliably. On the 163 posts where the two annotators agree (i.e. a clear human majority exists; the remaining 13 are genuine disagreements), the distilled labels reach 0.902 agreement with that ma-

Table 10: No-claim rate by MR2 class (LLM-distilled labels on the merged evaluation corpus). The unverified (label-2) class carries a no-claim rate $8.47\times$ that of the real class, and $8.5\times$ the real/rumor average.

Class	n	no-claim posts	no-claim rate
label 0 (real)	797	45	5.65%
label 1 (rumor)	650	46	7.08%
label 2 (unverified)	477	257	**53.88%**
real/rumor average	—	—	6.37%

jority ($\kappa = 0.804$). The key reading for construct validity is the gap between the two: the LLM judge's agreement with the human majority ($\kappa = 0.804$) is only slightly below the human agreement ceiling (inter-annotator $\kappa = 0.850$). Against each annotator individually over all 176 posts the judge reaches 0.869 (κ=0.740) and 0.875 (κ=0.751) respectively, again close to how well the two annotators agree with each other. The distilled labels are therefore not a judge idiosyncrasy: they reproduce a notion of verifiability that independent humans reproduce among themselves at nearly the same rate, which is what grounds the claim that the direction (trained on these labels) decodes a human-reproducible property rather than a judge-specific artifact. Table 11 summarizes the agreement statistics, including the per-domain breakdown.

Sampling uncertainty and two qualifications. Because the audit is small, we attach bootstrap 95% CIs (10,000 resamples of the audited posts): inter-annotator $\kappa = 0.850$ [0.76, 0.92] and distilled-vs-majority $\kappa = 0.804$ [0.71, 0.89]. The two intervals overlap substantially, so the "slightly below the ceiling" reading is within sampling noise rather than a demonstrated deficit. Two qualifications are worth stating explicitly. *(i) The audit tests the rubric, not nature.* Annotators were given the same rubric as the judge (Section 4.1), so a high κ says humans can apply this rubric consistently; it does not show that the rubric externalizes a natural kind, nor rule out a bias shared by rubric, judge, and annotators. An independent-rubric audit (a single-sentence definition, no categories) and an architecturally distinct second judge are the two natural next steps and are left to future work. *(ii) Agreement is asymmetric across classes.* On the consensus subset the judge matches annotator 1 on 0.947 of claim-bearing posts but only 0.812 of no-claim posts (Table 12): the disagreement concentrates on the class the gate depends on, so the no-claim labels are noisier than the headline κ suggests. The stratified design over-samples no-claim and boundary posts, which makes the overall agreement a *harder* test than a natural-prevalence draw but means the figure is not an unbiased corpus estimate.

The audit's only structural asymmetry is in the judge and does not affect the gate's training labels: `satire_parody` posts are treated by the judge as no-claim only when they carry no checkable proposition, whereas the *downstream router* sends them to forced binary adjudication for a different reason (satire mimics news and is still worth a $0/1$ verdict; Appendix P). The two decisions use different fields, `has_verifiable_claim` for the gate and `content_type` for routing, so no-claim labels are not contaminated by routing policy; we state this explicitly because the rubric table lists both fields side by side.

Independent-rubric re-audit. Qualification (i) above, that the audit tests the rubric rather than nature, is the review's central concern, so we ran the requested control: the same 176 posts were re-judged with a deliberately *different* instrument, a single-sentence plain-language definition with no category menu, no field list, and no access to the original rubric or its labels (same model family as the paraphrase generator, text+image input, verdict parsed from a forced `Y/N`; 173 of 176 returned a parseable verdict). The result is mixed and we report both halves. The independent instrument does *not* reproduce the original rubric's labels: against the human majority it agrees on 0.675 of posts ($\kappa = 0.374$, bootstrap 95% CI [0.25, 0.50]), well below the original rubric judge's $\kappa = 0.804$, and its no-claim base rate is 32.9% against 56.6% for the humans and 48.0% for the original judge. The disagreement concentrates where the construct is hardest: it matches the majority on 0.887 of claim-bearing posts but only 0.506 of no-claim posts; the one-sentence wording absorbs jokes and memes that *reference* checkable real-world events into the checkable class. Read against the over-sampled design (the audit deliberately concentrates boundary cases), the honest conclusion is

Table 11: Audit summary: agreement statistics (n=176, stratified by domain $\times$ distilled label $\times$ detector correctness; two annotators, Cohen's κ throughout).

Quantity	Value
Inter-annotator agreement (raw / κ)	0.926 / 0.850
AMG (n=60)	0.933 / 0.861
FineFake (n=60)	0.917 / 0.832
Weibo (n=56)	0.929 / 0.857
Distilled vs. human majority (raw / κ, n=163)	0.902 / 0.804
Distilled vs. annotator 1 (raw / κ, n=176)	0.869 / 0.740
Distilled vs. annotator 2 (raw / κ, n=176)	0.875 / 0.751

Table 12: Class-conditional agreement of the distilled judge with a human labeler on the audit sample (57.4% no-claim, far above the 10.8–13.7% natural rate). Agreement is markedly lower on the no-claim class the gate depends on, so the headline κ over-states label quality there.

Human reference	class	n	agreement
Annotator 1	no-claim	101	0.812
	claim	75	0.947
Annotator 2	no-claim	96	0.833
	claim	80	0.925
Human majority (consensus subset)	no-claim	92	0.859
	claim	71	0.958

that the *boundary* of the construct is instrument-sensitive, and we weaken the claim accordingly: we do not read the audit as showing that the rubric externalizes a natural kind, only that it names a human-reproducible convention. What the control does establish is that the *direction* is not one instrument's artifact: scored against the independent labels the probe still reaches AUROC 0.846 (AMG) and 0.834 (FineFake), against 0.971/0.977 against the original rubric on the same posts (the Weibo subsample is omitted because the audit's Weibo indices cannot be aligned to the local representation dump). A full *gate-level* re-evaluation under the independent rubric would additionally require re-distilling the training and validation corpora with it; we report the probe-level control here and leave that re-distillation to future work. We release the independent-rubric labels with the code.

Prior qualitative evidence. Before the formal audit, a 20-post stratified sample of FineFake rumor-bucket true positives labeled no-claim contained 15 Reddit PsBattle entries (edited images carrying no factual claim) and 5 claim-free short captions, 100% consistent with the distilled label; a 5-post false-positive contrast sample consisted entirely of opinions and jokes. Separately, human review of detector errors (with web-search access) found that most unverifiable posts cannot be adjudicated even in principle (Section 3).

H View Ablation

View selection uses the same protocol as the main probe (train fit, validation grid over $k \in \{32, 64, 128, 256\}$ and $\lambda \in \{0.003, 0.01, 0.03, 0.1\}$; each row is the best combination for that view on validation).

Two robust observations, plus a selection audit. First, modality separation dominates depth: at the same final layer, `txt_mean` exceeds the full-sequence mean by 13.8 AUROC points, so image tokens are a dilution source, and text-side pooling is itself a strong design decision; the image-only view never exceeds 0.79 in any configuration. Second, `last_token` is a close second in-domain, indicating that the classification head also reads the direction; under domain transfer it is markedly less stable across detector seeds than `txt_mean`, which is why the main protocol fixes the validation-selected view `txt_mean` rather than re-selecting per experiment (per-seed values in Appendix J).

Table 13: View ablation on AMG (validation AUROC). The text-token mean is the strongest carrier; the image-token view is barely above chance.

View	best (k, λ)	val AUROC
txt_mean (selected)	(256, 0.01)	**0.9344**
last_token	(256, 0.03)	0.9213
q3_mean (L27, full sequence)	(32, 0.1)	0.8076
mid_mean (L18, full sequence)	(128, 0.01)	0.7745
mean_last (L36, full sequence)	(256, 0.03)	0.7695
img_mean	(32, 0.003)	0.7693

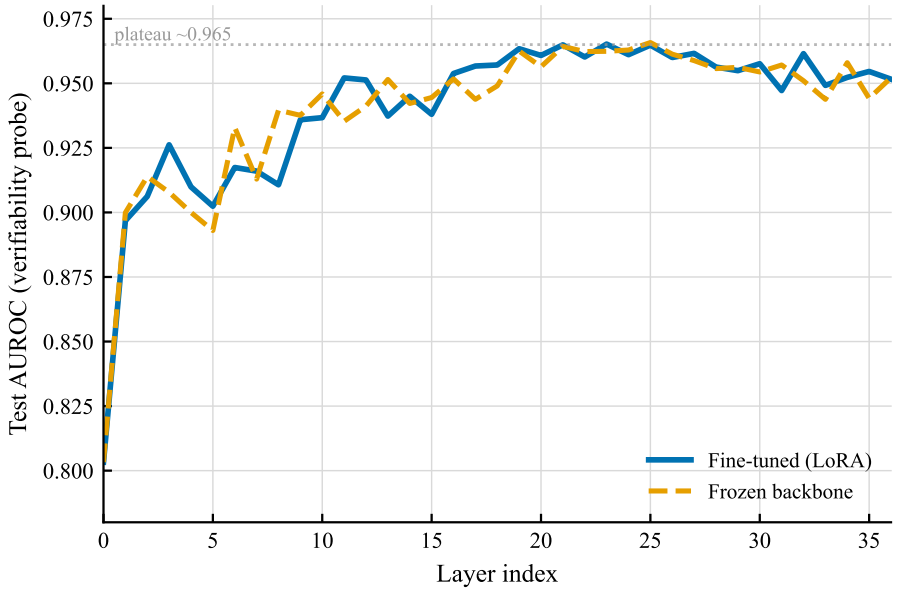


Figure 5: Verifiability AUROC of a per-layer probe across all 37 layers of Qwen3-VL-4B, for the fine-tuned detector and the frozen backbone. The two curves are nearly identical (mean $|\Delta| = 0.007$): on this backbone the direction emerges early and is a property of the pretrained representation space.

Circularity of a text-side construct. Because the no-claim label is defined on the text (`has_verifiable_claim`), a weak image-token view is partly *definitional*: our rubric does not label image-side unverifiability (a mismatched or manipulated image attached to a verifiable caption falls outside the no-claim target). The modality result therefore says the construct is carried by the claim text *under this rubric*, not that verifiability is independent of the image; we deliberately do not read it as an image-side finding. An image-aware re-distillation, that is, relabelling with a rubric that scores the image–text pair, is the natural test and is left to future work.

Selection audit. View and hyperparameters are chosen on validation while the test split is scored once, so we report the whole selection surface rather than the winning cell. Over all 96 (view, k, λ) combinations the test AUROC spans 0.47–0.96, median 0.83; within `txt_mean` alone it spans 0.49–0.95. The validation-selected configuration's test AUROC (0.9512) ranks 3rd of 96 and sits 0.005 below the grid maximum (0.956, a `last_token` cell), so the headline figure is not a lucky cell. The low end comes from under-regularized collapses at k=256; the good region is a broad plateau of text-side configurations, and the test-side ordering of the text-side views is stable, which is what the reported numbers rest on.

Figure 5 shows the per-layer emergence analysis on the 37-layer Qwen3-VL-4B detector: the direction is present from the first layer, plateaus around layers 23–25 slightly above the final layer, and the frozen-backbone curve is indistinguishable from the fine-tuned one (mean $|\Delta| = 0.007$).

Table 14: In-domain head (val-fit) vs. the AMG transfer head (seed 42, the fixed-`txt_mean` head of Table 15), test AUROC.

Domain	in-domain head	AMG head (seed 42)	difference
FineFake	0.9712	0.9834	transfer $\geq$ in-domain
Weibo	**0.9780**	0.9614	in-domain exceeds transfer

Table 15: Zero-shot transfer of the verifiability direction (AMG-trained head, fixed `txt_mean` view, three detector seeds, mean $\pm$ std).

Target domain (language)	n	no-claim rate	test AUROC
AMG, source (EN)	871	13.2%	0.9473 ± 0.0054
FineFake (EN)	1682	10.8%	$\mathbf{0.9827 \pm 0.0007}$
Weibo (ZH)	937	13.7%	0.9583 ± 0.0033

I IN-DOMAIN VS. TRANSFERRED HEAD

FineFake and Weibo lack training annotations for no-claim, so in-domain heads are fit on their *validation* sets (1683 and 940 posts) with 5-fold cross-validation for view/hyperparameter selection (the test split is never read during selection) and evaluated once on test, against the AMG head applied zero-shot. Both domains independently select `txt_mean` + PCA-256 on data disjoint from AMG and from each other, which is itself evidence that the carrier is not a property of the source domain.

Two practical conclusions: (i) on FineFake, zero-shot transfer matches or beats small-sample in-domain training, as the direction is consistent with a source-side data-volume effect rather than domain match; (ii) on Weibo, 940 in-domain labels recover the full transfer loss and slightly exceed it. For a new domain, the recommended deployment is the AMG transfer head cold (no annotation), refined with a few hundred local labels when they become available.

J CROSS-MODEL FAMILIES

Table 15 reports the source row and the zero-shot transfer of the verifiability direction; Table 16 lists the per-backbone details behind Section 4.1. Each backbone is fine-tuned as its own detector with the same LoRA recipe, its own distilled labels (same judge pipeline), and its own probe (view and PCA dimension selected on validation per backbone). Cross-model values are seed-42 single runs.

K SHALLOW-TEXT CONTROL: FULL DETAILS

Table 17 summarizes the control against the probe; the paragraphs below give the full protocol.

Protocol. The control differs from the main probe in exactly one respect: the input features are TF-IDF over the post text instead of pooled MLLM representations. Everything else is held fixed: the aligned text-bearing sample set (AMG val n=456, test n=802), the train/val/test splits, the distilled no-claim labels, the λ grid $\{0.003, 0.01, 0.03, 0.1\}$, validation-based model selection, and a single test evaluation. Feature families are word 1–2-gram, char_wb 3–5-gram, and their union, each with optional SVD to $k \in \{32, 64, 128, 256\}$ and the same ℓ_2-regularized logistic regression solver as the probe; 108 configurations are searched on AMG validation. The selected configuration is char_wb 3–5-gram, SVD k=128, λ=0.1.

Feature-family breakdown (AMG). Each family selects its own best (k, λ) on validation. Test AUROC: word 1–2-gram 0.827 (k=256, λ=0.003), char_wb 3–5-gram 0.860 (k=128, λ=0.1), union 0.850 (k=256, λ=0.003). Character n-grams dominate: the signal that survives is partly sub-word, consistent with a partial lexical basis.

Table 16: The verifiability direction across four MLLM families (each backbone fine-tuned as its own detector; probe protocol identical). Detector accuracy and all test AUROC values are per-backbone seed-42 single runs. LLaVA-OneVision-0.5B is named for its 0.5B language model; the total parameter count including the vision tower is 0.9B.

Backbone (family)	Params	detector acc	best view	test AUROC
Qwen3-VL-4B	4.4B	0.844	txt_mean	**0.9512**
Qwen2.5-VL-3B	3.7B	0.824	last_token	0.9401
SmolVLM2-2.2B	2.2B	0.840	txt_mean	0.9336
LLaVA-OneVision-0.5B	0.9B	0.786	last_token	0.8886

Table 17: Shallow-text control (TF-IDF + ℓ_2 LR) vs. the probe, identical protocol and aligned text-bearing subsets (n=802 AMG test; probe values recomputed on the same subset, within 0.004 of the full-set values in Table 15). "Refit" rows retrain the shallow model on target-domain labels (its supervised upper bound); the probe remains the zero-shot AMG head.

Protocol	Domain	shallow text	probe
in-domain fit	AMG	0.860	**0.9512**
zero-shot transfer	FineFake (EN)	0.930	**0.9827**
zero-shot transfer	Weibo (ZH)	0.530	**0.9583**
refit on target labels	FineFake	0.975	**0.9827**
refit on target labels	Weibo	0.925	**0.9583**

Wide-grid robustness. To rule out a regularization-limited baseline, we extend the grid to $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and drop the SVD constraint. The best wide-grid configuration (char_wb 3–5-gram, no SVD, λ=10^{-4}) reaches test AUROC 0.848, 0.012 below the narrow-grid winner. The baseline is not under-regularized.

In-domain upper bounds on target domains. FineFake and Weibo carry no train-split distilled labels, so their in-domain shallow models are fit on validation with 5-fold CV selection (both domains independently select TF-IDF + SVD k=128, λ=0.003; Weibo tokenization via jieba). Test AUROC: FineFake 0.975 (word+char union), Weibo 0.925 (jieba word 1–2-gram). These supervised upper bounds are the "refit" rows of Table 17; both remain below the zero-shot probe.

Why cross-lingual transfer collapses. The AMG-trained vocabulary covers 51.3%/49.0% of Weibo val/test tokens; the lexical model transfers essentially nothing (0.521/0.530, near chance). The gap to the probe (0.961) is structural: no amount of additional *English* lexical supervision can close it, since the failure is the vocabulary itself. The probe's representation space, by contrast, encodes verifiability in a language-shared geometry; the cross-lingual result is the strongest evidence that the direction is a representational property rather than a lexical one.

Concatenation probe. [TF-IDF-SVD ‖ repr] (char_wb 3–5-gram, k=256, λ=0.003) reaches test AUROC 0.949 on AMG, against 0.947 for the representation alone and 0.860 for TF-IDF alone. The increment of lexical features over the representation is +0.002: the representation already subsumes the lexical signal relevant to verifiability.

L Comparison with Published Methods on the Same Benchmarks

This appendix also collects the two cross-system views deferred from Section 2: the generational cost ladder (Figure 6) and the per-system cost/accuracy table (Table 19).

Table 20 positions our results against published systems on the four benchmarks we use. Four protocol caveats bound the reading, and we state them before the numbers.

First, the task formulation differs on AMG. Published AMG numbers (82.43, 86.10, etc.) are on the official multi-class *attribution* task over the official test split (n=973); our detector is the binary real/fake merge on our own split (n=874). Merging manipulation subtypes removes confusions

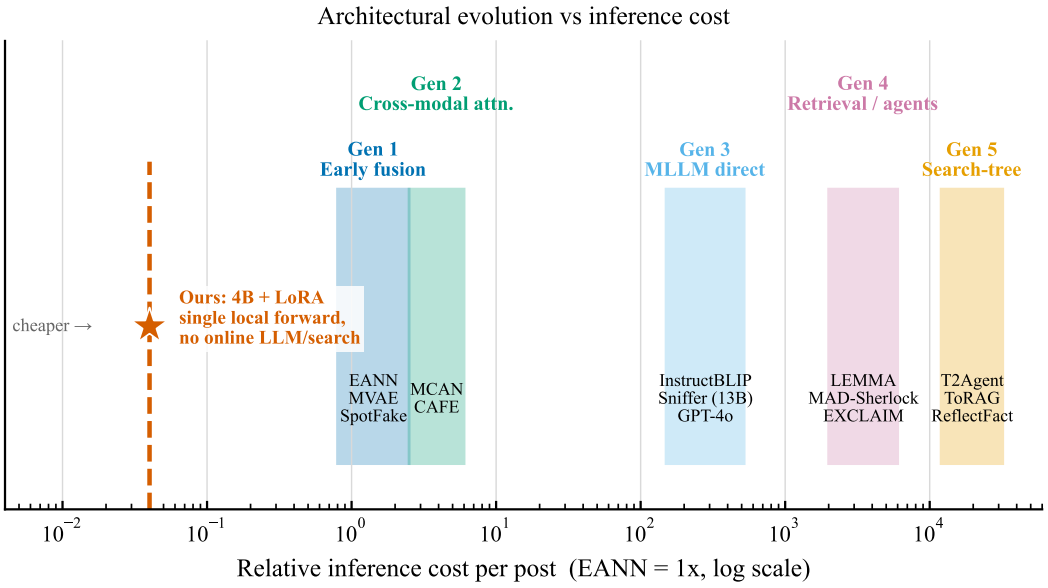


Figure 6: Architectural generations of multimodal rumor detection and their per-post inference cost (relative to early-fusion EANN = $1\times$, log-x). Agentic (generation 4–5) systems cost 10^3–$10^4\times$ more; a 4B MLLM used as a plain classifier, our substrate, costs strictly *less* than the generation-1 anchor (single local forward pass, zero online LLM calls or retrieval). The 4/5 boundary within agentic systems follows each source paper's own description of its pipeline (e.g. T2Agent plans with MCTS over retrieved evidence) and is indicative only; the cost claim rests on the aggregate 4–5 band.

Table 18: The baseline detector (frozen Qwen3-VL-4B + LoRA + linear head) under original splits. AMG is reported as mean over three seeds (42/43/44); precision/recall/F1 and the Weibo row are seed-42 values.

Dataset	test Acc	Precision	Recall	F1
AMG (3 seeds)	0.862 ± 0.017	0.721	0.970	0.827
FineFake	0.809	0.790	0.878	0.832
Weibo	0.961	0.950	0.969	0.960

the attribution task must resolve, which is why our binary figure sits above the published numbers there. We therefore read the AMG column as a protocol-mismatched positioning anchor rather than a comparison, and we do not claim competitiveness on the attribution task.

Second, the splits and corpora differ. On FineFake and Weibo, published systems use the official corpora and splits (FineFake 16,909 posts over six domains; classic Weibo), while we train and evaluate on our own splits; our numbers are positioning anchors, not leaderboard entries.

Third, the evidence setting differs on MR2. Published MR2 systems consume the provided external evidence pool, since retrieval-augmented pipelines are the point of that benchmark (TERN, Holmes, AAR). Our MR2 configuration is deliberately self-contained: no online retrieval, no online LLM, evaluated on the merged test set plus out-of-scope posts with the tiered handler resolving rejections. It answers a different deployment question (what can be refused and handled offline) and is not aiming at the retrieval-augmented operating point.

Finally, no published method has an abstention mechanism. Every number in the upper blocks is answered-accuracy over forced predictions; none of these systems can be evaluated under the risk–coverage protocol that our evaluation centers on, because none can decline to answer. This is precisely the capability gap our contribution targets.

Table 19: Reported binary real-vs-fake accuracy (%) on each dataset and per-post cost across published multimodal rumor detectors (2018–2026). Each accuracy cell is on the *column's* dataset so rows are comparable only *within* a column; "–" means the paper reports no binary accuracy there. Cost is per-post (online API + search, or amortized local GPU); each row reports the numbers stated in its original publication, and the classical detectors are cited in Section 2; single-forward discriminative models and 13B local MLLMs both sit in the 10^{-4}–10^{-3} band. Our row reports the measured detector cost of Appendix Q (single local forward pass, below the band); Weibo21 is not in our benchmarks.

System	Architecture	USD/post	AMG	FineFake	Weibo	Weibo21
T2Agent (GPT-4o)	MCTS agent	1.64	78.40	–	–	–
MMD-agent (GPT-4o)	agent pipeline	0.34	71.50	–	–	–
EXCLAIM	3-agent pipeline	0.02	–	–	–	–
MGCA	reverse-image search	0.008	80.89	–	–	–
SNIFFER	13B + search	0.009	–	–	–	–
TRUST-VL	7–13B local	0.002	–	–	–	–
MOSAIC	∼220M discrim.	0.0002	–	80.60	–	–
GC2MFND	∼180M discrim.	0.0001	–	81.20	95.30	95.70
CAFE	∼150M cross-modal	0.0001	75.85	77.40	84.00	88.30
MCAN	cross-attention	< 0.001	74.61	76.20	88.50	89.60
EANN	early fusion	< 0.001	73.18	74.10	78.80	82.30
MVAE	early fusion	< 0.001	71.94	72.50	82.40	84.50
InstructBLIP	13B, zero-shot	< 0.01	66.80	68.30	81.50	–
Ours	4B + LoRA + probe	1.7×10^{-5}	**86.2**	**80.9**	**96.1**	–

With those caveats registered, on Weibo the deliberately minimal recipe (0.961) is competitive with, and above most of, the published specialized detectors, below only RoViT's 98.93, which we do not explain away (RoViT uses modality gating over the full corpus, and our split differs); on FineFake it trails the published binary systems by 2–3 points (84.30/82.51), consistent with those systems' use of the full six-domain corpus and explicit knowledge alignment; on AMG the binary merge (0.862) is not comparable to the published numbers because of the task and split difference noted above; and on MR2 the full system (0.812) is not comparable to retrieval-augmented accuracies by design. We therefore make no state-of-the-art claim on *any* benchmark: the paper's claim is about the rejection layer these systems do not have, and Table 20 exists to show that the accuracy axis is contested enough that accuracy alone cannot be the contribution.

FineFake cross-domain adaptation (PANDA, 74.12%) and the Weibo-21 multi-domain benchmark (MDFEND, 94.30%) address different task settings (unsupervised domain transfer; a different corpus) and are omitted.

M Statistical Significance

Routing improvements (Table 1). For each domain the two-gate rule routes a subset of test samples to rejection, and we report the effect *directionally* (the false positives each policy intercepts and the answered-set accuracy) rather than as a paired significance table. A per-sample paired McNemar over the detector's binary output is not the applicable test here, for a structural rather than a sample-size reason: the two-gate policy is a *deferral* mechanism. On the samples it rejects, the detector does not emit a 0/1 label that can be scored as right or wrong against confidence-only's guess: the sample is handed to the second-level handler of Section 4.3, and a post the handler declines to answer is left as an *abstention* (an unanswered post, not a misclassification). The two policies therefore disagree on decision space, not merely on a set of labels, so there is no shared 0/1 label whose discordant pairs form a clean paired test of a single intervention. The recovery of the rejected remainder is itself an empirical quantity, reported in Section 4.3: the resolver issues a judgment on each rejected post (falling back to the detector where necessary), and the fully-answered end-to-end accuracy is not below the detector's direct accuracy (Section 4.3). Under that protocol we read the intercept of additional false positives as a consistent directional gain across domains rather than an individually significant cell; the Weibo rumor-side bucket is also the smallest (n=23), which only reinforces the decision-space argument over a per-cell test.

Table 20: Published results on the benchmarks we use, positioned against ours. Upper rows per block: published systems under their own protocols (task, split, and evidence setting as indicated); bottom row: ours under our protocol. "RAG" marks systems that consume an external evidence pool. No method in the published blocks has an abstention mechanism.

Benchmark (task)	Method	Acc (%)	F1 (%)	Setting
AMG (attribution task, official split)	ExDR (SIGIR 2026)	86.10	84.60	5-LVLM RAG average
	MGCA (AAAI 2025)	82.43	81.56	multi-granularity alignment
	CAFE (Chen et al., 2022) (WWW 2023)	78.62	77.45	ambiguity-aware fusion
	MCAN (Wu et al., 2021) (ACL 2022)	75.18	74.30	co-attention
	T2Agent (Cui et al., 2026) (AAAI 2026)	72.80	71.90	training-free MCTS agent
	Ours (detector)	**86.2±1.7**	82.7	binary merge, our split, n=874
FineFake (binary)	RAMM (arXiv 2026)	84.30	83.70	LLM narrative alignment
	KEAN (FineFake'24)	82.51	81.93	knowledge-graph align
	DMGMoE (ICIC 2024)	81.80	80.95	multi-granularity MoE
	Ours (detector)	**80.9**	83.2	our split, n=1688
Weibo (binary)	RoViT (2024)	98.93	98.10	modality gating
	ERIC-FND (AAAI 2025)	93.80	93.80	contrastive interaction
	AAR (Zheng et al., 2025) (COLING 2025)	93.10	91.40	MLLM debate distill
	MMCAN (2022)	91.60	91.90	co-attention
	COOLANT (ACM MM 2023)	91.10	90.80	contrastive alignment
	CAFE (Chen et al., 2022) (WWW 2023)	90.80	90.80	ambiguity-aware fusion
	Ours (detector)	**96.1**	96.0	our split, n=942
MR2 (three-way)	TERN (MDPI 2025)	93.21	91.09	RAG (provided pool)
	AAR (Zheng et al., 2025) (COLING 2025)	91.30	90.70	RAG, EN subset
	Holmes (2024)	88.30	87.60	open retrieval
	SparseMoE-MFN (2024)	86.70	85.90	sparse MoE
	Ours (full system)	**81.2**	—	self-contained, no RAG, merged + OOS, n=1308

What *is* shared between the two policies is the rejection event itself: a detector false positive is either intercepted or answered by each policy, which defines a common binary outcome on the common FP set and permits paired inference without forcing the policies into a common label space. At the frozen validation-selected operating points we compute, per domain, (i) the 2×2 interception table over the FP set with an exact McNemar test; (ii) a *nested variant* that pins the margin threshold to the confidence-only arm's τ (the two-gate rejection set then strictly contains the confidence-only set, the extra interceptions come entirely from the verifiability gate, and the two-sided sign test applies); and (iii) paired bootstrap 95% CIs ($10{,}000$ resamples of the test set) for the FP-interception and answered-accuracy differences (Table 21; reconstruction verified against the frozen per-domain reports before testing). The extra interceptions are significant in all three domains under the nested variant ($p \leq 9.8 \times 10^{-4}$) and under the resampling CIs; at the matched operating points the exact test gives $p = .057$ on AMG, conservative since the two arms' val-tuned margin thresholds differ and three FPs are intercepted by the margin-only arm alone; it gives $p = .0014$ on FineFake and $p = 3.1 \times 10^{-5}$ on Weibo. The answered-accuracy differences are small and their CIs straddle zero (AMG [-0.002,+0.016], Weibo [-0.003,+0.017]) except for the FineFake trade already disclosed in Section 4.2 ([-0.037,-0.014]): the gate's value is interception and rumor-side precision at matched coverage, not answered accuracy.

System-level gaps (Table 2). Differences between the full system and each ablation are assessed with a bootstrap over the merged test set ($10{,}000$ resamples, percentile CIs):

For the two ablation rows (no-evidence, self-contained gating) the CI is computed over the *paired per-sample predictions within a matched detector directory* (the LLM-distilled gate vs. its probe replacement run under identical conditions), i.e. a strict single-variable ablation of the gate. The self-contained row's point gap under this matched comparison (Δ=0.035) is slightly smaller than the cross-directory gap read off Table 2 (0.812 vs. 0.758), because the full-system row there uses the DO-LoRA detector variant while the ablation pair shares a single detector; the bootstrap interval is the honest estimate for the gate-only effect.

Table 21: Paired significance tests for the routing improvements of Table 1 (frozen validation-selected thresholds; test scored once). "Int." = FPs intercepted; discordant pairs = intercepted by the confidence-only arm only / by the two-gate arm only; nested = margin threshold pinned to the confidence-only arm's τ (two-sided sign test on the gate's extra interceptions); CIs from 10,000 test-set resamples.

Domain	FP set	Int. conf→gate	Discordant	McNemar p	Nested p	ΔInt. 95% CI
AMG	126	78→86	3/11	.057	9.8×10^{-4}	$[+1, +16]$
FineFake	211	186→204	6/24	.0014	1.2×10^{-7}	$[+7, +29]$
Weibo	23	6→22	0/16	3.1×10^{-5}	2.4×10^{-4}	$[+9, +24]$

Table 22: Bootstrap 95% CIs for the accuracy gaps of the full system over the ablations on MR2.

Comparison	95% CI of Δaccuracy
full system − no-evidence ablation	[0.006, 0.031]
full system − self-contained gating ablation	[0.022, 0.049]

Risk–coverage comparisons (Table 5). At the matched ≈95%-coverage points, the two-gate rule's +2.8-point answered-accuracy gain carries a paired bootstrap 95% CI of [+1.8,+3.7] (10,000 resamples of the merged set). The label-2 leakage reduction ($492 \rightarrow 441$ of 506) is a shared-event comparison on the label-2 posts and is significant under an exact McNemar ($p = 9.1 \times 10^{-10}$). At the lowest confidence-only-reachable coverage (79%), the confidence-only gate still leaks 428 of 506 label-2 posts (84.6%, Wilson 95% CI [81.2%,87.5%]).

The two-domain extension is consistent: the verifiability-blocking arm likewise fails to beat the unblocked baseline (seed 42; random control on AMG only): 0.8092→0.7891 on FineFake, 0.9607→0.9533 on Weibo. The direction should therefore be read as a *marker* that tracks the failure, robustly enough to gate rejection on, rather than the mechanism that produces it; this is why our method reads the direction out at inference (where it demonstrably helps) rather than manipulating it at training time, and why we report the control rather than a cherry-picked seed.

For seed stability and the blocking experiment, the AMG detector's test accuracy is stable across the three detector seeds (42/43/44: 0.8444–0.8787; mean 0.8619 ± 0.0172). Table 23 reports the three-arm blocking comparison as seed-level accuracies and their mean±std; we do not report a three-seed paired t-test between the arms because the per-sample predictions for seeds 43 and 44 of the DO-LoRA and random arms were not retained on disk, so a sample-level paired test over all three seeds is not available from the stored artifacts.

Cost side of the gate: how many rejected posts were answerable? The routing tables report intercepted false positives and answered accuracy, but neither is a cost measure. We therefore report, for every frozen operating point, the composition of the rejected set: the number of rejected posts the detector would in fact have answered *correctly*. Table 24 gives the decomposition. On AMG the verifiability gate adds 44 rejections over confidence-only; those buy 8 additional intercepted FPs but discard 36 posts that would have been answered correctly, so the +0.6-point answered-accuracy gain is partly a denominator effect (fewer, easier remaining posts) rather than a genuine accuracy improvement. On FineFake and Weibo the effect is much larger: 74.7% and 94.1% of the rejected posts would have been answered correctly. This is the price of a precision-targeted, scope-aware operating point, and it is why we present the frontier (Figure 3) rather than a single cell, and why reject-only is the honest alternative deployment mode (Section 4.3).

Small-sample caution on Weibo. The Weibo rumor-side bucket contains only 23 false positives, and the interspersion test is an *exact* McNemar (the discordant split is $0/16$, so the two-sided $p = 3.1 \times 10^{-5}$ equals the exact binomial sign probability) rather than an asymptotic χ^2 approximation; the paired bootstrap CI for the interception difference is [+9,+24]. An exact test on a bucket this small is still fragile, so we treat the Weibo cell as a directional illustration of the sign criterion, corroborated by the larger FineFake and AMG buckets and by the mechanism-level rejection-source statistics (Appendix B), not as an independently established effect. On AMG the matched-point

Table 23: Training-time subspace blocking on AMG (three detector seeds).

	seed 42	seed 43	seed 44	mean ± std
baseline	0.8444	0.8627	0.8787	0.8619 ± 0.0172
block verifiability dir.	0.8867	0.8330	0.8627	0.8608 ± 0.0270
block random dir.	0.8787	0.8799	0.8799	$\mathbf{0.8795 \pm 0.0007}$

Table 24: Composition of the rejected set at each frozen operating point. "Would-be-correct" counts rejected posts the detector classifies correctly; the last column is that count as a share of all posts the detector gets right. AMG confidence-only uses the margin-only threshold.

Domain / gate	coverage	answered acc	rejected	would-be-correct	share of correct
AMG, confidence-only	87.1%	0.924	113	35	4.7%
AMG, + verifiability	82.0%	0.930	157	71	9.6%
FineFake, + verifiability	52.2%	0.866	807	603	44.1%
Weibo, + verifiability	60.4%	0.974	373	351	38.8%

exact McNemar is $p = .057$ and is reported as non-significant; the significant figure there ($p = 9.8 \times 10^{-4}$) belongs to the nested variant with the margin threshold pinned, which isolates the verifiability gate's incremental interceptions and is a different question. Whether the Weibo trade is worth taking is a deployment decision rather than an empirical one: the gate exchanges $\sim$39 coverage points for 16 additional intercepted FPs and $+0.026$ rumor-bucket precision (0.963$\rightarrow$0.989), which favours the gate when a single published rumour is expensive and disfavours it when throughput binds; we therefore report the frontier rather than a single recommended point.

N Direction Control, Non-Posterior Baselines, and Trivial-Fill Bounds

This appendix reports three controls added in revision: what happens if the gate direction is *not* selected per domain (Section 4.2), whether training-free input-side heuristics recover the no-claim signal, and how much of the fully-answered cascade's margin over the detector is available to a trivial fallback. All numbers are recomputed from the frozen per-sample dumps and reproduce the published operating points bit-for-bit.

Forcing one direction across domains. The sign criterion sets the gate direction from a single validation statistic, and a natural worry is that a criterion validated on three domains is fragile. Table 25 evaluates the counterfactual directly: at each domain's frozen operating point we hold τ and κ fixed and flip *only* the direction. The criterion is load-bearing on FineFake, where the AMG direction collapses the gate (7 of 211 FPs intercepted and rumor-bucket precision 0.747, far below the 0.95 target the operating point is tuned for) and second-order on Weibo, where the forward direction remains a usable gate (18 of 23) but concedes four interceptions and 22 coverage points. On AMG the flipped direction intercepts more FPs but at a precision (0.731) that also violates the target. Forcing the forward direction everywhere is therefore not a viable uniform setting, and on FineFake no forward κ reaches the reverse arm's interception without collapsing coverage (the best forward cell intercepts all 211 FPs at 40.3% coverage). We read the criterion not as a theorem but as a per-domain operating-point selector whose failure mode is measurable and bounded; the direction is a genuine degree of freedom of the gate, not a notational detail.

Training-free input-side heuristics. The baselines of Section 4.2 are posterior-derived; a reviewer asks whether a *non-posterior* signal (surface cues such as questions, first person, exclamations, and message length) could do the job without any probe. We score nine such cues (question marks and interrogatives, first-person markers, exclamation density, capitalisation ratio, short-text indicator, URL presence, digit presence, emoji presence, log length) on the same evaluation sets as the probe. No single cue exceeds 0.76 AUROC on AMG test, and a logistic stack over all nine, fit on AMG train and hence using strictly more supervised information than the probe's zero-shot transfer setting, reaches 0.742 on AMG test (0.710 refit on AMG validation) against the probe's 0.9512. On

Table 25: Direction control at the frozen operating points (validation-selected τ, κ; test scored once; only the gate direction is flipped). Cells report coverage / FPs intercepted; the selected direction per domain is bold. Flipping the direction is not interchangeable: it collapses the FineFake gate and violates the precision target on AMG.

Domain	confidence only	forward ($p_{nc} < \kappa$)	reverse ($p_{nc} \geq \kappa$)
AMG	87.1% / 78	**82.0% / 86**	54.2% / 112
FineFake	74.3% / 186	88.1% / 7	**52.2% / 204**
Weibo	99.3% / 6	82.9% / 18	**60.4% / 22**

Table 26: Fully-answered accuracy when the rejected set is filled trivially instead of routed to the handler (same frozen operating points; detector accuracy on the answered set unchanged). "Always rumor" recovers the detector exactly, since the gate only rejects rumor-side predictions.

Domain	rejected	rumor share	always real	random	cascade (reported)
AMG	157	45.2%	86.16%	85.30%	86.04%
FineFake	807	74.7%	57.29%	69.11%	83.20%
Weibo	373	94.1%	61.15%	78.61%	96.44%

MR2 the stack reaches 0.875/0.826 on val/test (0.839 refit on MR2 validation) against the probe's 0.914. The cues carry part of the signal, since short posts are indeed more often no-claim, but the increment the representation adds is 0.11–0.21 AUROC, and the cue-based stack is also the least stable across domains, which is what a surface-statistics proxy should look like.

How much of the cascade's margin is available to a trivial fallback? Because the gate only abstains on rumor-side predictions, the rejected set is rumor-heavy (45.2%/74.7%/94.1% rumor on AMG/FineFake/Weibo), which puts a low ceiling under any trivial fill. Table 26 reports the fully-answered accuracy obtained by filling the rejected set with constant or random labels instead of running the handler. On FineFake and Weibo the reported cascade beats *every* trivial fill ($+2.3$ and $+0.4$ points over the strongest), so the handler earns its cost there. On AMG the cascade sits 0.1 points *below* the strongest trivial fill (answering every rejected post "real" would have done slightly better), which is the concrete reason we present reject-only as an equally honest deployment on that domain and price the handler separately (Appendix Q). The cascade's value is domain-conditional, and the bound here makes that condition explicit rather than leaving it implicit in a coverage–accuracy curve.

O Scope of the Confidence Baseline: Full Argument

This appendix expands the structural argument of Section 4.2 for the baselines we subsume under "confidence-only" and the one family we exclude.

Any signal derived from the detector's output distribution over its trained labels (MSP, margin, temperature-scaled scores, conformal wrappers on those scores, MC-dropout variance, or multi-seed checkpoint ensemble disagreement) estimates the model's posterior spread over *labels*. Conformal wrappers and temperature scaling are monotone in the margin, which is why they produce bit-identical routing decisions in our setting (Appendix A); MC-dropout and ensembles spread the posterior more broadly but rank the same posts, by construction: an unverifiable post yields a confidently wrong label under any posterior, and no amount of posterior spreading exposes it. The failure mode we target is a property of the *input*, namely whether the post belongs to the modeled label space at all, so no output-side label-uncertainty signal can see it.

Semantic-consistency estimators (Kuhn et al., 2023) are excluded for a structural reason rather than an empirical one. Semantic entropy requires free-form generation whose semantic clustering captures meaning-level diversity; our detector is a discriminative head over a fixed label set, so sampled outputs *are* labels and output-side semantic consistency degenerates exactly to the label spread already measured by MC-dropout and ensembles. The non-degenerate variant is input-side con-

Table 27: The four regimes carved out by the two near-orthogonal signals (margin m, no-claim posterior p_{nc}).

	high margin m	low margin m
verifiable ($p_{nc} < \kappa$)	answered directly; most accurate slice	knowledge failure; retrieval can settle it; second stage earns its cost
not verifiable ($p_{nc} \geq \kappa$)	*dangerous cell*: confident on nothing to check; sign criterion: refuse or "rumor"	nothing to verify; handler often declines; judgment falls back to the detector (Section 4.3)

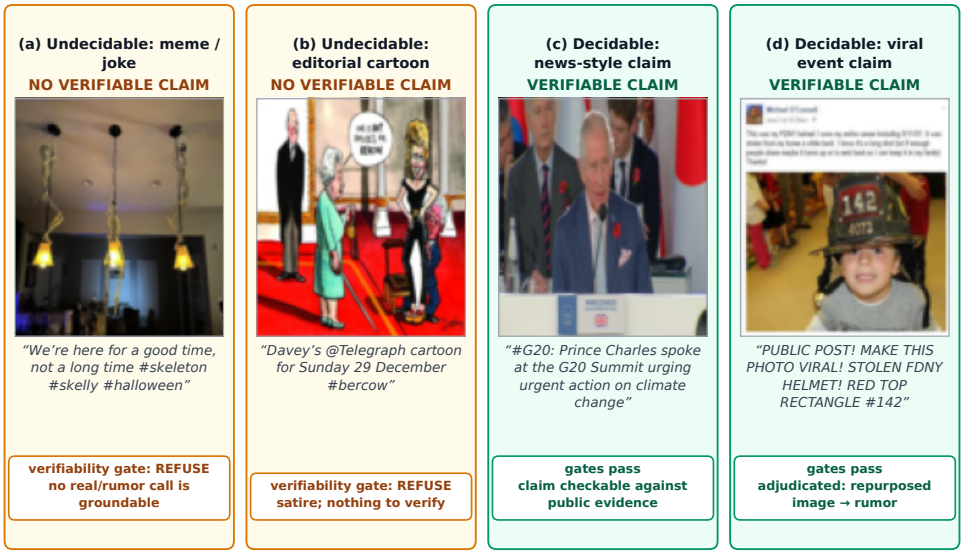


Figure 7: Real AMG posts illustrating the two axes (images from the public benchmark). (a,b) Posts asserting no verifiable claim (memes, jokes, cartoons): any real/rumor output is ungroundable, so gate 2 refuses them regardless of confidence. (c,d) Claim-bearing posts: a judgment is possible in principle, by checking the claim against public evidence. The verdict strips show the routing policy given verifiability, not per-post measurements.

sistency under paraphrase of the multimodal post: it perturbs the input rather than the model, and probes robustness rather than verifiability. We therefore ran it.

Input-side consistency, measured. For each merged MR2 post we paraphrase the text 8× with an instruction-tuned LLM (meaning, claims, and language preserved; image unchanged; prompt in Appendix P); one post is excluded because the generation model's content policy refuses it. The detector runs on all rewrites and each post is scored by the agreement rate between its paraphrase predictions and the original prediction. The result is an *inversion*, not a signal: the most paraphrase-stable content in the corpus is precisely the out-of-scope class (mean agreement 0.99 for label-2 vs. 0.96 for ordinary posts; 92.7% of all posts are perfectly stable, and the few fully unstable posts are ordinary hard cases, none of them label-2). At its strictest threshold the consistency gate still answers 92.7% of posts with 488/506 label-2 leaks, and answered accuracy *falls* slightly (0.560→0.558); matched at ≈95% coverage it sits at or below the confidence-only baseline it augments (answered accuracy 0.559 vs. 0.5611; leakage 496 vs. 492), at nine detector forwards per post instead of one. This is the empirical counterpart of the structural argument: paraphrase consistency re-measures *difficulty*, and difficulty is not the axis on which this failure lives.

P PROMPTS AND RUBRICS

All prompts below are the exact prompts run online (verbatim from code, not paraphrased). The judge and agent are run with the same English prompts across all domains, including Weibo, regardless of the post's language (Appendix G). The choice buys a single judge implementation, and therefore comparable label semantics, across all domains; a Chinese-prompt control run reproduces the same results, so prompt language is not a live confound. The audit of Appendix G bounds the downstream risk of the judge design, since its annotators are bilingual for Weibo and English-capable elsewhere, and inter-annotator agreement on Weibo ($\kappa = 0.857$, Table 11) matches the English domains. The judge, the MR2 handler, and the paraphrase generator are served through the SenseNova chat-completions endpoint, and the handler's retrieval backend is the Bocha web-search API; exact model identifiers, decoding parameters, and per-invocation prices are part of the released configuration. We give the prompts' structure, JSON fields, and judgment rules; the line references point to the released code.

The verifiability judge (no-claim distillation). Source: `annotate_guidance.py`, `GuidanceLabeler.PROMPT_TEMPLATE` (with image) and `PROMPT_TEMPLATE_NO_IMAGE` (text-only). The prompt frames the task as *structural analysis of a claim*: it instructs the model to analyze the "claim structure" of a post, and *explicitly forbids* any true/false judgment: it never sees the label, relation, claim, or target, and is told that deciding truth is "not your task." Its output is a JSON object with the following fields (definitions authoritative; see rubric below):

- `visual_summary`: one-sentence objective description of the image (empty string when no image).
- `image_text_relation`: one of `support`/`irrelevant`/`contradict`/`no_image`. `support` = the image provides direct visual evidence for the text's claim; `irrelevant` = no substantive relation (decorative, diagram, portrait); `contradict` = visible contradiction; `no_image` = no image.
- `relation_evidence`: one or two sentences justifying the relation judgment.
- `content_type`: the content category (see the rubric table below).
- `has_verifiable_claim`: boolean, whether the text contains a concrete claim that can be checked against external evidence (news reports, official records, public data). This is the operationalization of the paper's no-claim label.
- `main_claim`: one-sentence restatement of the central verifiable claim, in the original language (empty string if none).
- `verification_focus`: the aspect most in need of verification: `text_facts`/`image_origin`/`image_text_consistency`/`source`/`none`.
- `verification_targets`: up to three concrete claim points to verify.
- `verification_questions`: up to three concrete queries ready to hand to a search engine.

The text-only variant fixes `visual_summary` to `""`, `image_text_relation` to `no_image`, `relation_evidence` to a placeholder, and narrows `verification_focus` to `text_facts`/`source`/`none`.

Rubric: `content_type` / `has_verifiable_claim` / `verification_focus`.

`content_type`	Meaning	is_news
`news_report`	News: report/alert/photo-news of a real-world public event	yes
`opinion_commentary`	Opinion/editorial/analysis of a public event	yes
`satire_parody`	Satirical fake news mimicking news style (not real reporting)	no*
`fan_community`	Fandom/derivative work; discussion of fiction	no
`personal_life`	Everyday sharing, emotional expression	no
`ad_promotion`	Advertising/recruitment/event notice	no
`meme_joke`	Meme/joke/banter	no
`other`	Other	no

*The judge's is_news derivation (annotate_guidance.py) admits only news_report/opinion_commentary; the agent's routing (below) additionally sends satire_parody to forced binary, because satire still warrants a 0/1 adjudication. has_verifiable_claim is boolean; no-claim = it is false. verification_focus is one of the five values above (text_facts/image_origin/image_text_consistency/source/none; text-only narrows to the first and the last two).

Three-way agent resolution (with evidence). Source: agent_mr2.py, LLMDecisionAgent._build_prompt (no_evidence=False). Used after the conformal/rule gate has rejected a post, to issue a final three-way decision (prediction $0/1/2$) from structured retrieval (image search, text search, image–text entity consistency, source-credibility summary). The judgment procedure is: (1) if the content is not in the news/public-information scope (fan work, fiction discussion, fandom, personal life, ads, pure opinion/meme), output 2 (non-news); (2) otherwise adjudicate truth: 0 if mainstream media report the same event with consistent image and text; 1 if the image is repurposed, the source is low-credibility or satirical, mainstream media do not report the claim, or the caption spreads false/misleading material; 2 only if the evidence is severely contradictory. Key principles: rumor often travels as short text/question/exclamation with an image, so short claims about real-world events still undergo step 2 and lean 1 absent mainstream backing; mainstream coverage of an *image* is not coverage of the *caption's claim*. The no-evidence ablation (no_evidence=True) removes all four evidence blocks and relies on text style alone, keeping the same non-news-first structure.

Forced-binary agent (news-category posts) and routing rules. Source: agent_mr2.py, _build_prompt_binary. Because news-category posts in the rejection bucket are rumor-dominant with almost no mainstream coverage, the "cannot decide" escape hatch is removed and the agent must output 0 or 1. Routing (lines 768–770) by the judge's content_type:

- NON_NEWS_TYPES = { fan_community, personal_life, ad_promotion, meme_joke } → answered as non-news by the rule layer (no binary LLM call).
- NEWS_TYPES = { news_report, opinion_commentary, satire_parody } → the forced-binary prompt below issues the final 0/1 decision.

The forced-binary prompt gives the content, the independent structural annotation (content_type, main_claim), and the four evidence blocks; it instructs: output 0 only if mainstream headlines/summaries narrate the same event *and* the image–text are consistent *and* the writing is objective with concrete specifics; output 1 on any of (no mainstream coverage, repurposed image, social/low-credibility-only source, sensationalist/vague sourcing such as "reportedly"/"netizens say"/unsourced "BREAKING", satirical content, or tampering with real specifics). It stresses that *absence of evidence is itself a signal*: a news claim with no mainstream coverage and no text evidence is likely rumor.

In-system decision protocol. The agent's decisions match the evaluation scripts: conf¡0.5 ⇒ pred=2 (abstention counts as label 2; evaluate_three_tasks.py, agent_final_pred/is_low_conf_abstain). The conformal gate rejects first; rejected posts are routed by the judge's content_type: NON_NEWS_TYPES to the rule layer, NEWS_TYPES to forced binary, and the remainder to the three-way prompt above.

Paraphrase generation (input-side consistency baseline, Appendix O). For each merged MR2 post, an instruction-tuned LLM receives the post text and the instruction: rewrite the post so that meaning, claims, stance, and level of specificity are identical; wording, sentence order, and phrasing change substantially; the output language matches the input; no claim is added, removed, or strengthened; return exactly 8 distinct rewrites as a JSON list (temperature 0.8). The image is never perturbed. One post is refused by the generation model's content policy and excluded from this baseline's curve.

Verbatim prompt templates. For reproducibility we reproduce below, verbatim as run online, the two prompts that carry the paper's decisions: the no-claim judge that defines the construct, and the

forced-binary resolver that issues the final $0/1$ verdict. Content inserted at run time is marked by angle-bracket placeholders, and line breaks are typesetting only. The evidence-conditioned three-way variant (`LLMDecisionAgent._build_prompt`, same file) uses the same four evidence blocks and the two-step procedure described above and is reproduced in the released code.

(a) No-claim judge. Source: `annotate_guidance.py`, `GuidanceLabeler.PROMPT_TEMPLATE` (image variant).

```
You are an expert in news image-text structure analysis. Analyze the
"claim structure" of the following social-media post objectively.

[IMPORTANT] Your task is structural analysis only. You are strictly
forbidden from judging the content true or false, or giving any truth
lean or probability -- that is not your task.

Text content:
<post text>

Output exactly one JSON object (and nothing else), with these fields:
{
  "visual_summary": "one-sentence objective description of the image
    (empty string if there is no image)",
  "image_text_relation": "relation of image to textual claim; must be
    one of support/irrelevant/contradict/no_image. support = the image
    provides direct visual evidence for the textual claim; irrelevant =
    the image bears no substantive relation to the textual claim (e.g.,
    decorative photo, illustration, portrait); contradict = the image and
    the textual claim visibly conflict; no_image = there is no image",
  "relation_evidence": "the concrete visual/textual basis for the
    relation above (one or two sentences)",
  "content_type": "content type; must be one of: news_report = news
    reporting (report/alert/photo-news of a real-world public event);
    opinion_commentary = opinion/editorial/analysis of a public event;
    satire_parody = satirical fake news from a parody site, mimicking
    news style but not real reporting; fan_community = fan/derivative
    works, celebrity fandom, discussion of fiction; personal_life =
    everyday sharing, emotional expression; ad_promotion = advertising/
    recruitment/event notice; meme_joke = meme/joke/banter; other",
  "has_verifiable_claim": true or false -- whether the text contains a
    concrete claim that can be checked against external evidence (news
    reports, official records, public data),
  "main_claim": "the single most central verifiable claim, one sentence
    (keep the original language; empty string if none)",
  "verification_focus": "the aspect most needing verification; must be
    one of text_facts/image_origin/image_text_consistency/source/none.
    text_facts = the textual facts themselves; image_origin = image
    provenance and time/place of capture; image_text_consistency =
    whether image and text describe the same event; source = source
    credibility; none = nothing to verify",
  "verification_targets": ["concrete factual points to verify, at most
    3, one sentence each"],
  "verification_questions": ["concrete queries ready to hand to a search
    engine/retrieval system, at most 3"]
}
```

(b) Forced-binary resolver. Source: `agent_mr2.py`, `LLMDecisionAgent._build_prompt_binary`.

```
You are an expert in news fact-checking. The classifier is uncertain
about the content below; you must issue the final verdict.

[CONTENT TO CHECK]
<post text>

[CONTENT STRUCTURE ANALYSIS] (from an independent annotation system;
structure only, no truth judgment)
```

```
- content type: <content_type>
- core claim: <main_claim>

[Evidence 1: image provenance (reverse image search)]
<image-search matches>
Stats: <n> exact matches, of which <n> from mainstream media.

[Evidence 2: text provenance (search by text)]
<text-search matches>
Stats: <n> text matches, of which <n> from mainstream media.

[Evidence 3: image-text entity consistency]
<entity overlap>

[Evidence 4: source-credibility summary]
Mainstream-media coverage present: <yes/no>

[YOUR TASK]
This content has been confirmed to be within the news/public-information
scope. Decide whether it is real news (0) or rumor (1).

[DECISION CRITERIA] (ordered by priority)
Output 0 (real news) only if ALL of the following hold:
  - mainstream-media headlines/summaries narrate the same event
  - image and text are consistent (high entity overlap)
  - the wording is objective and the specifics are concrete (time,
    place, people, official sources)

Output 1 (rumor) if ANY of the following holds:
  - no mainstream media reports the claim (rumors precisely lack
    mainstream backing and spread via social media)
  - the image is repurposed (image-text mismatch, entities do not
    match, the same image paired with different claims)
  - the source is only social media or low-credibility sites
  - the text style is sensationalist and the sourcing is vague
    ("reportedly", "netizens say", unsourced "BREAKING")
  - satirical parody from a joke site (mimics news style but is not
    real reporting)
  - concrete details of a real event have been tampered with or
    exaggerated

[IMPORTANT] You must output 0 or 1; outputting 2 is forbidden. The
absence of evidence is itself a signal: news-type claim + no mainstream
coverage + no text evidence -> most likely rumor (1).

Output exactly one JSON object:
{
  "prediction": 0 or 1,
  "confidence": a float from 0.0 to 1.0,
  "reasoning": "one or two sentences explaining the decision"
}
```

Q PER-POST COST MEASUREMENT

Table 28 reports self-measured inference cost for our own pipeline. It complements the cross-system comparison of Table 19, whose published rows each use their source paper's own accounting (API list prices, amortized GPU, or reported budget); those rows are therefore indicative rather than strictly commensurable with a single measurement protocol, and we keep the two tables separate for that reason.

Measurement basis. Throughput and GPU cost are measured for the detector alone: frozen Qwen3-VL-4B with LoRA and the linear head, one batched forward pass per post, on a single rented GPU at 1.4 CNY/hour sustaining 12,000 posts/hour, that is, 0.30 s of GPU time per post,

Table 28: Self-measured per-post inference cost. Relative cost is normalized to the detector alone. GPU seconds are measured; online-LLM and retrieval calls are structural for our pipeline (zero for detector and probe, one per rejected post for the handler) and published for the agentic row (Table 19). AMG is reported at its directly-measured 0.95-precision rejection rate; FineFake and Weibo are reported as the range across the MR2-measured rejection-rate interval 8%–20%, because their throughput rejection rate is not directly measured and their Table 1 operating point is an upper envelope, not a deployment rate. Call counts are expectations per post, not integers.

Configuration	GPU s/post	online LLM	retrieval	CNY/post	relative
Detector (4B, single forward)	0.30	0	0	1.2×10^{-4}	$1\times$
+ verifiability probe	0.30	0	0	1.2×10^{-4}	$\approx 1\times$
+ handler, AMG (18.0% rejected)	0.30	0.18	0.18	6.6×10^{-3}	$57\times$
+ handler, FineFake (8%–20% rejected)	0.30	0.08–0.20	0.08–0.20	3.0×10^{-3}–7.3×10^{-3}	25–61×
+ handler, Weibo (8%–20% rejected)	0.30	0.08–0.20	0.08–0.20	3.0×10^{-3}–7.3×10^{-3}	25–61×
Agentic pipeline (T2Agent, published)	–	5–80	5–80	11.00	$9.4 \times 10^4\times$

or 1.2×10^{-4} CNY (1.7×10^{-5} USD, converted at the market rate on the measurement date, $1\,\text{USD} = 6.71\,\text{CNY}$, 2026-09-06). The verifiability probe adds no measurable cost: it is a PCA projection followed by a logistic-regression dot product, applied to hidden states the detector has already computed, so it requires no additional forward pass and no additional model. The second-level handler is priced at 0.036 CNY per invocation (one retrieval-augmented agent judgment), invoked only on posts the gates reject, so it amortizes as (rejection rate) $\times$ 0.036 plus the detector cost.

The rejection rate is the quantity that sets the end-to-end cost, and it depends on operating point. On AMG the gate's coverage 82.0% is measured directly at the 0.95-precision operating point (Table 1), so the end-to-end row is a single, reproducible number: 18.0% of posts rejected, all at one handler invocation apiece. We do *not* carry FineFake and Weibo to a single point, because the coverage we report for them in Table 1 is a deliberately conservative 0.95-precision operating point that rejects far more than ordinary content flow would; it is an upper envelope, not a throughput estimate. For those two domains we instead show the end-to-end cost over the rejection-rate range 8%–20%, spanning the rates directly measured on MR2 (7.7% of test and 19.3% of validation posts; Appendix B) as a realistic low-to-mid deployment reference. Because the detector dominates only at very low rejection rates, the detector cost in Table 28 includes both the base 1.2×10^{-4} CNY forward pass and, in the full rows, the handler; the end-to-end rows are stated as ranges whose endpoints are the detector alone plus the handler cost at 8% and 20% rejection respectively.

What the numbers say. Two facts carry the deployment argument. First, *the second level dominates cost*: at the directly measured AMG operating point, adding the handler multiplies per-post cost by 57×, and even across the moderate rejection range 8%–20% the multiplier reaches ∼61×, so the detector is nearly free by comparison. The economically decisive component is therefore not the backbone but the gate, which decides which posts incur that multiplier, exactly the object this paper studies. Second, even at the top of the rejection range the end-to-end cost ($\sim 7.3 \times 10^{-3}$ CNY $\approx \sim 1.1 \times 10^{-3}$ USD per post) stays 18×–1500× below the agentic pipelines of Table 19, and the detector alone is $9.4 \times 10^4\times$ cheaper than T2Agent. The *Ours* marker in Table 19 is the measured detector cost (1.7×10^{-5} USD), which is what the accuracy experiments use; the gated pipeline that resolves rejected posts is only invoked when a platform chooses the cascade and is at most a few $\times 10^{-3}$ USD even in the upper part of the range.

Caveats. These are measurements of one implementation on one hardware configuration at one rental price; absolute figures move with hardware, batch size and provider, and the *ratios* are the robust part. The detector figure is the steady-state rate of the serving loop (image decoding, the single batched forward pass, and the linear head); it excludes the handler's request pattern, and because rejected posts leave the detector path, batch composition under a high rejection rate is not modelled. The handler is priced as a single agent invocation per rejected post; a deployment issuing multiple retrieval rounds, or a multi-agent debate, per post scales the handler term linearly in the number of invocations and leaves the detector and probe terms unchanged. Finally, we measure cost, not latency or energy; wall-clock behaviour under production batching is not characterized here.