

Generation-Score Consistency: Pushing LLMs to Generate Semantically Valid Reasoning Scores for Misinformation Detection

Yidu Chen^{1,2,3} Bo Ma^{1,2,3*} Yating Yang^{1,2,3*} Xi Zhou^{1,2,3}
Zhen Wang^{1,2,3} Ahtamjan Ahmat^{1,2,3} Dilxat Abdureyim^{1,2,3} Azmat Anwar^{1,2,3}

¹Xinjiang Technical Institute of Physics & Chemistry,
Chinese Academy of Sciences, Urumqi, China

²University of Chinese Academy of Sciences, Beijing, China

³Xinjiang Laboratory of Minority Speech and Language
Information Processing, Urumqi, China

chenyidu22@mailsucas.ac.cn

{mabo, yangyt, zhousi, wang_zhen, ahtmj, dilxat, azmat}@ms.xjb.ac.cn

Abstract

Large language models (LLMs) possess extensive world knowledge and have demonstrated strong potential for misinformation detection. A growing line of recent work prompts LLMs to produce structured reasoning before issuing a veracity judgment. However, we contend that such structured outputs do not spontaneously support the veracity decision. The emitted scores can satisfy the output schema while contributing little to veracity discrimination. Since annotating full reasoning paths is prohibitively costly for veracity tasks, we propose a Generation-Score Consistency framework that supervises through structured dimension scores and steers the LLM toward functional reasoning. Specifically, we train a compact score-to-label verifier on a small set of expert-annotated reasoning data, freeze it, and during supervised fine-tuning use a trainable score head to align the score-token hidden states of the LLM with the human-calibrated mapping. Introducing structured reasoning on its own costs accuracy relative to label-only supervision, while the consistency term recovers that cost and keeps the emitted scores coupled to the veracity decision. Controlled experiments on multimodal misinformation benchmarks show that the consistency term strengthens both veracity prediction and the discriminability of the emitted scores. Code and data are available at <https://github.com/yun-nlp/Reasoning-Scores-for-Misinformation-Detection-EMNLP2026>.

1 Introduction

Misinformation spreads rapidly across social media and affects high-stakes domains such as healthcare, politics, and crisis response, making scalable

veracity detection an important task for modern information systems (Lewandowsky et al., 2012; Osmundsen et al., 2021; Nielsen and McConville, 2022; Yousuf, 2025; Lao et al., 2024).

With the rapid progress of large language models (LLMs), a common practice is to prompt LLMs to produce structured analyses of multiple aspects, dimension scores, and stepwise rationales alongside a final veracity judgment (Wu et al., 2024; Wan et al., 2024; Hu et al., 2024; Park et al., 2025). However, outputs that follow the required format do not establish that these intermediate judgments support the decision. When the gold label is included in the supervision sequence, an LLM can satisfy the output schema while leaving its scores semantically inert. We refer to this gap as the difference between *structured* and *functional* reasoning, whether the emitted scores actually contribute to the veracity decision or merely accompany it. At the same time, directly annotating reasoning paths at scale is prohibitively costly, since veracity assessment is a holistic judgment that experts cannot easily decompose into stepwise chains.

To address the challenges outlined above, we propose Generation-Score Consistency (GSC), a lightweight mechanism that uses structured dimension scores as a compact form of expert supervision and binds scores generated by the LLM to the veracity decision through a frozen verifier calibrated using human annotations. We train a compact verifier g_ϕ to map score vectors to veracity labels using 600 posts rated by experts along six information quality dimensions, then freeze it as a prior on which score patterns indicate real or fake content. During supervised fine tuning, a trainable score head f_ψ projects the hidden states at the positions of the six score tokens into a score vector for this verifier. A consistency loss requires the prediction

* Corresponding author.

of the verifier to agree with the gold veracity label. Unlike methods that use verifiers to assess reasoning chains expressed in free text (Lightman et al., 2024), GSC operates over a compact score space interpretable by humans, limiting human supervision to a small calibration set.

Does the gain of structured-reasoning SFT come from the semantics of the emitted scores, or are the structured outputs merely superficial? We therefore evaluate GSC with a family of variants that hold the gold label and output schema fixed, and perturb only the score signal, the verifier, or the readout path; on top of this, we conduct hidden-state interventions at the score-token positions to test whether those representations are behaviorally active during generation. Our contributions are summarized as follows:

- We identify and formalize the gap between structured and functional reasoning in LLM-based misinformation detection, and argue that current evaluation protocols overlook it.
- We propose GSC, a low-supervision consistency mechanism that uses a frozen, human-calibrated verifier to bind LLM-generated reasoning scores to veracity, as an early attempt at verifier-consistent reasoning scores for fake news detection without dense annotation.
- Through label- and format-matched controls, score-token positional probes, and hidden-state interventions, we attribute the gain specifically to score semantics and to behaviorally active score-token representations.

2 Related Work

2.1 LLM-based Misinformation Detection

Multimodal misinformation detection has progressed from text-only classifiers to systems that fuse textual, visual, and social-context signals (Nielsen and McConville, 2022; Hu et al., 2025), and increasingly uses LLMs as reasoning or augmentation modules through dimension-style critiques (Wan et al., 2024), rationale-augmented classifiers (Hu et al., 2024), in-context demonstrations (Xu et al., 2024), and generated comments, rewrites, or agent reactions (Nan et al., 2024; Park et al., 2025; Wu et al., 2024; Cheung and Lam, 2023); a parallel line uses LLMs to summarize retrieved evidence for human fact-checkers (Roitero et al., 2025). These methods

place LLM-produced signals beside the final label without binding them to the LLM’s own decision, so structured rationales or scores can stay decorative while the decision forms elsewhere, echoing broader concerns that chains of thought can diverge from the computation behind a model’s decision (Turpin et al., 2023). GSC targets this gap by checking the LLM’s generated reasoning scores against a frozen, human-calibrated verifier that the model cannot retune during SFT.

2.2 Verifier-based Reasoning and Information Quality

Verifier-based reasoning trains auxiliary models to score generated solutions for reranking or supervision (Lightman et al., 2024; Yu et al., 2024; Hosseini et al., 2024), and self-consistency further suggests that agreement between intermediate reasoning and the final answer is itself informative (Wang et al., 2023). We adapt this idea to misinformation detection by verifying a compact score vector, which keeps the check interpretable and bounds human supervision to a small calibration set. This score vector is grounded in information-quality research that decomposes fitness-for-use into dimensions (Stvilia et al., 2007; Wand and Wang, 1996; Juran, 1992; International Organization for Standardization, 2008), validated for credibility judgments by experts and crowd workers (Soprano et al., 2021; Maddalena et al., 2018; Barbera et al., 2024; Ceolin et al., 2023; Visser et al., 2020), where assessor background and bias are known to affect the resulting scores (Draws et al., 2022) and motivate our use of domain experts for the calibration set. We use these six dimensions as a structural carrier for LLM reasoning, and leave the mapping from scores to veracity to the frozen verifier (§3).

3 Information Quality Dimensions and Human Annotation

GSC needs a compact space in which an LLM can state intermediate judgments and an external verifier can check whether those judgments support the final label. We instantiate this space with six information-quality dimensions grounded in prior multidimensional assessment work (Soprano et al., 2021; International Organization for Standardization, 2008), namely **Neutrality**, **Correctness**, **Completeness**, **Precision**, **Comprehensibility**, and **Informativeness**. These dimensions provide a human-interpretable carrier for structured

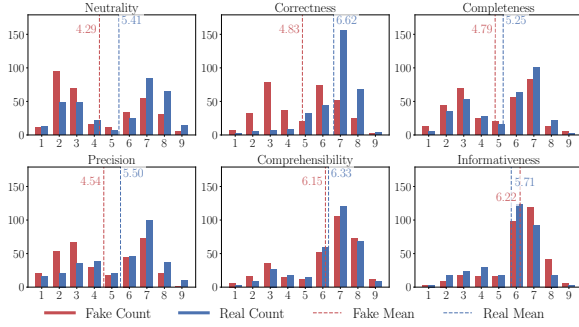


Figure 1: Human score distributions across the six information-quality dimensions. Each dimension carries label-related structure that the verifier combines into a veracity judgment.

reasoning scores that can be supervised and verified, while the verifier decides how they combine into a veracity judgment.

We collect a small calibration set $\mathcal{D}_{\text{human}}$ of 600 posts, balanced across Weibo-21 (Nan et al., 2021) and Twitter (Boididou et al., 2018), and ask five expert annotators to rate each post on the six dimensions using a 1-to-9 scale. These annotations are used only to calibrate the preference-aligned generator and the frozen score-to-label verifier g_ϕ ; no human scores are collected for the full training corpus. Figure 1 shows that several dimensions separate real and fake posts with partial overlap, which motivates using them as a verifier-consistent reasoning structure that feeds the label decision. Full definitions, annotation protocol, agreement, and distributional evidence are provided in Appendix B.1–B.2.

4 Generation-Score Consistency

Our goal is to make an LLM learn a structured, human-readable reasoning path that produces the final veracity label for fake-news detection. The six information-quality scores provide this intermediate structure, but human-scoring the full corpus is costly, and naive SFT on generated score-and-label outputs can still let the model learn the label and format while treating scores as decorative text. GSC handles these obstacles in sequence, as summarised in Figure 2. First, a small human-scored set aligns a generator with the scoring rubric, so the model learns both the output schema and the human preference over score profiles. The aligned generator then scales this supervision to the training corpus under hidden-label sampling. However, generated structured targets still contain the gold

veracity label during SFT, so the final model may satisfy the sequence loss by learning the label token while leaving the scores weakly used. We therefore add a frozen verifier g_ϕ , which encodes the human-calibrated score-label relation, and a trainable score head f_ψ , which connects LLM score-token representations to that verifier. The resulting objective requires the generated score profile to agree with the gold veracity label under a verifier the LLM cannot retune.

4.1 Human Alignment and Large-Scale Score Annotation

The six dimensions introduced in Section 3 define the score space that the model must learn to use. Let $\mathcal{D}_{\text{human}}$ denote the 600 expert-annotated posts, each paired with a six-dimensional score vector and a binary veracity label. We first align the backbone LLM to this rubric with Group Relative Policy Optimization (GRPO), because the generator used for later annotation must produce score profiles that follow human scoring preferences on top of a valid output schema. The GRPO reward combines format validity, veracity correctness, and a group-relative preference that ranks the sampled completions by how close their scores are to the expert scores, yielding a generator M_{grpo} whose score profiles follow expert scoring preferences; reward details and hyperparameters are deferred to Appendix C.1 and Table 10. Because scoring the full training corpus by hand would defeat the low-supervision goal, we use M_{grpo} to annotate the rest of the data.

Concretely, M_{grpo} builds the structured supervision corpus $\mathcal{D}_{\text{LLM}}$ under hidden-label sampling, where the prompt hides the gold label and the generator must infer a reasoning trace, six scores, and a veracity prediction from the post. This prevents the annotation model from directly copying the label into its reasoning trajectory. A sampled candidate is accepted as a *strict success* only when its predicted label matches the gold label and all structured score and reason fields parse cleanly. The first strict-success candidate becomes the SFT target $\tilde{o}(x)$; if no candidate succeeds, the instance falls back to a label-only target. The binary mask m_x records this distinction, with $m_x = 1$ for strict-success instances that carry scores and $m_x = 0$ for fallback instances whose consistency loss must be disabled.

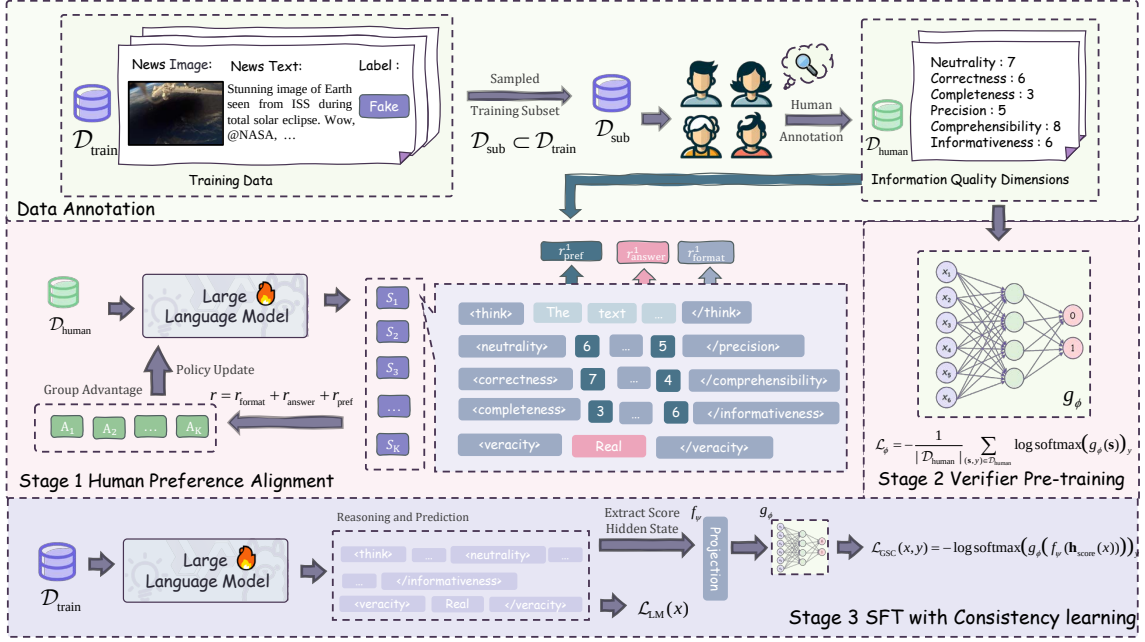


Figure 2: Three-stage pipeline of GSC. Stage 1, Human Preference Alignment, aligns the backbone with the human scoring rubric using GRPO and uses the aligned generator to annotate the training set under hidden-label sampling. Stage 2, Verifier Pre-training, trains a compact score-to-label verifier g_ϕ on expert scores and freezes it as the human-calibrated reference. Stage 3, SFT with Consistency Learning, trains the final LLM with language modeling plus a gated consistency loss, where score-token hidden states are projected by f_ψ and checked by the frozen g_ϕ . Blue components are trainable; the gray verifier is frozen.

4.2 Consistency Module: Frozen Verifier and Score Head

Even with $\mathcal{D}_{\text{LLM}}$ in hand, naive SFT on these structured targets cannot ensure that the generated scores carry veracity-relevant semantics. A model can still copy the schema, optimize the final label token, and leave the six scores weakly coupled to the decision. We therefore add a consistency module that makes the score profile itself consistent with an expert-calibrated score-to-label relation.

To do this lightly, we encode the score-to-label relation in a compact human-calibrated verifier g_ϕ and freeze it, so the LLM cannot satisfy the consistency objective by co-adapting the checker. Concretely, g_ϕ is trained on $\mathcal{D}_{\text{human}}$ to map a six-dimensional score vector to two-way veracity logits, then frozen for the rest of training.

Given an input x and its target output sequence $\tilde{o}(x)$ from $\mathcal{D}_{\text{LLM}}$, the LLM is run in teacher-forcing mode, and we locate the six score-token positions in the structured output. Let $\mathbf{h}_i \in \mathbb{R}^d$ denote the SFT readout hidden state at the i -th score-token position; in the main objective this is the last-layer hidden state, and stacking the six positions yields

$$\mathbf{h}_{\text{score}} = [\mathbf{h}_1, \dots, \mathbf{h}_6] \in \mathbb{R}^{6 \times d}. \quad (1)$$

Feeding hidden states directly into g_ϕ poses a representational mismatch, since g_ϕ is calibrated on six numeric scores while LLM hidden states are high-dimensional representations shaped by the language-modeling objective. Forcing g_ϕ onto this mismatched space would put the language-modeling and verifier objectives in direct competition for the same parameters, making the auxiliary objective brittle and expensive to train. We therefore introduce a trainable score head f_ψ as an adaptor, which reads exactly the score-token positions and projects them into a continuous six-dimensional score estimate,

$$\hat{\mathbf{s}} = f_\psi(\mathbf{h}_{\text{score}}) \in \mathbb{R}^6. \quad (2)$$

Architectural details are in Appendix A.1. The resulting estimate is checked by the frozen verifier,

$$\hat{\mathbf{z}} = g_\phi(\hat{\mathbf{s}}) \in \mathbb{R}^2, \quad (3)$$

where $\hat{\mathbf{z}}$ are the verifier’s logits for the LLM-generated score profile.

4.3 SFT Training Loss

With g_ϕ fixed and f_ψ providing a differentiable score readout, final SFT can combine ordinary sequence learning with score-label consistency. The

consistency loss requires the verifier’s verdict to match the gold veracity $y \in \{0, 1\}$, where 1 denotes fake and 0 denotes real:

$$\mathcal{L}_{\text{GSC}} = \text{CrossEntropy}(\hat{\mathbf{z}}, y). \quad (4)$$

Thus $\mathcal{L}_{\text{GSC}}$ penalizes score profiles that look valid as text but do not imply the correct label under the human-calibrated verifier.

The final training objective combines the standard language-modeling loss with the consistency term, gated by the annotation-time mask m_x :

$$\mathcal{L}_{\text{total}}(x) = \mathcal{L}_{\text{LM}}(x) + m_x \beta \mathcal{L}_{\text{GSC}}(x). \quad (5)$$

Here $\mathcal{L}_{\text{LM}}$ is next-token cross-entropy over the target sequence and β controls the verifier constraint. The mask is necessary because fallback instances have no generated score fields, so strict-success instances train both terms while fallback instances train only $\mathcal{L}_{\text{LM}}$. We use $\beta = 0.7$ as the main setting on both Weibo21 and Twitter, with the full β sweep reported in Appendix A.2. Because strict-success targets include both generated scores and the gold veracity token, a model could still learn the label and format without meaningful scores; our controlled variants in Section 5.3 therefore keep $\mathcal{D}_{\text{LLM}}$, the prompt, and gold-label tokens fixed, changing only the score signal supplied to g_ϕ or the dimensions available to it.

5 Experiments

We organize the experiments around the mechanism behind the gain. After setup and main comparison context, we isolate verifier-side score semantics, probe the score-head readout path, intervene on score-position hidden states, inspect human-calibrated score behavior, and close with a qualitative sanity check.

5.1 Setup

Datasets. The Weibo21 dataset (Nan et al., 2021) contains 4,640 real and 4,488 fake news instances. We follow the standard split of 4,176 real and 4,039 fake instances for training, and report results on the Weibo21 test split ($n=913$). The Twitter dataset (Boididou et al., 2018) consists of 5,819 real and 7,547 fake instances for training and a Twitter test split of $n=1,091$.

Training Configuration. We adopt **Qwen2.5-VL-7B** (Bai et al., 2025) as the backbone for all variants and use the same LoRA/SFT recipe for final

training. GRPO preference alignment uses a separate LoRA setting; full GRPO/SFT hyperparameters, prompts, and module details are listed in Appendix D and Table 10. The main consistency weight is $\beta = 0.7$ for both datasets; all positive- β Weibo21 controls use the same value, and the full β sweep is deferred to Appendix A.2.

Baselines. We compare GSC with multimodal detectors and LLM-based methods listed in Table 1; Appendix E.1 summarizes each baseline. These baselines provide operating context, while our main evidence comes from controlled GSC variants that hold the backbone, label supervision, output schema, and training recipe fixed.

Evaluation Protocol. All classification metrics are recomputed from model outputs using 1=fake and 0=real, and Macro-F1 is the unweighted mean of real-news F1 and fake-news F1. For paired mechanism comparisons, we align samples by instance index and report aggregate metrics, paired accuracy gaps, 10k-bootstrap 95% confidence intervals, two-sided McNemar tests, and error-flip counts.

5.2 Main Comparison Results

The most direct comparison in Table 1 is against Qwen2.5-VL-7B variants under the same backbone family. On Weibo21, GSC reaches 95.0% accuracy with balanced real/fake F1 near 95% (Table 1). The comparison isolates a clear pattern. Ordinary SFT can already exploit task labels, and adding score consistency leaves the operating point essentially unchanged while making the emitted scores functional, without shifting performance toward only real or fake news. On Twitter, GSC reaches 89.7% accuracy (Table 1), while specialized multimodal detectors remain strong. Among parsed outputs, GPT-5.2 tends to over-predict fake under our zero-shot prompt, lowering real-news recall; unparseable outputs are treated as fake in metric computation, which affects GPT-4o most strongly on Twitter. Table 1 therefore establishes the operating context for the controlled studies that follow. The stronger evidence for our claim comes from paired mechanism studies where label, schema, backbone, and training recipe are controlled.

5.3 Ablation Study

Table 2 collects the ablation results. All controls hold the backbone, prompt, output schema, training steps, and gold veracity label fixed, changing only the score signal or the verifier configuration. **Reasoning-only** removes the verifier constraint by

Dataset	Method	Accuracy	Real News			Fake News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	EANN (Wang et al., 2018)	0.648	0.584	0.759	0.660	0.810	0.498	0.617
	SAFE (Zhou et al., 2020)	0.762	0.695	0.811	0.748	0.831	0.724	0.774
	SpotFake (Singhal et al., 2019)	0.785	0.792	0.675	0.729	0.780	0.867	0.821
	CAFE (Chen et al., 2022)	0.806	0.805	0.813	0.809	0.807	0.799	0.803
	BMR (Ying et al., 2023)	0.878	0.890	0.816	0.852	0.870	0.925	0.896
	MCAN (Zhang et al., 2025)	0.917	0.992	0.854	0.918	0.852	0.992	0.917
	BMLHF (Wu et al., 2025)	0.966	0.948	0.975	0.956	0.985	0.933	0.957
	GPT-3.5-turbo	0.568	0.498	0.690	0.578	0.672	0.477	0.558
	GPT-4o	0.686	0.610	0.739	0.669	0.767	0.645	0.701
	GPT-5.2	0.613	0.565	0.429	0.488	0.637	0.751	0.689
	SuperICL (Xu et al., 2024)	0.844	0.853	0.769	0.809	0.839	0.900	0.868
	ARG (Hu et al., 2024)	0.853	0.860	0.786	0.821	0.849	0.904	0.876
	Qwen2.5-VL-7B (zero-shot)	0.556	0.485	0.560	0.520	0.626	0.554	0.588
	GSC (ours)	0.897	0.830	0.957	0.889	0.964	0.852	0.905
Weibo21	EANN (Wang et al., 2018)	0.870	0.841	0.912	0.875	0.902	0.825	0.862
	SpotFake (Singhal et al., 2019)	0.851	0.786	0.964	0.866	0.953	0.733	0.828
	CAFE (Chen et al., 2022)	0.882	0.907	0.844	0.876	0.857	0.915	0.885
	BMR (Ying et al., 2023)	0.923	0.941	0.905	0.923	0.906	0.942	0.924
	DAPT (Tong et al., 2025)	0.951	—	—	0.950	—	—	0.952
	GPT-3.5-turbo	0.686	0.652	0.817	0.725	0.744	0.550	0.633
	GPT-4o	0.736	0.782	0.666	0.719	0.701	0.808	0.751
	GPT-5.2	0.740	0.935	0.526	0.673	0.663	0.962	0.785
	SuperICL (Xu et al., 2024)	0.873	0.914	0.828	0.869	0.837	0.920	0.877
	ARG (Hu et al., 2024)	0.897	0.919	0.875	0.896	0.877	0.920	0.898
	Qwen2.5-VL-7B (zero-shot)	0.660	0.689	0.606	0.644	0.638	0.717	0.675
	GSC (ours)	0.950	0.954	0.946	0.950	0.945	0.953	0.949

Table 1: Main comparison results on Twitter and Weibo21. Methods above the dashed line are pretrained-model based; methods below it are large language model based.

setting $\beta = 0$. **Random-G** and **Shuffled-G** keep the same supervised target and label but corrupt the scores seen by the consistency path, either with valid random score vectors or with a fixed cross-instance permutation. **Trainable-G** lets g_ϕ co-adapt with the LLM, and **Reduced-Dim** restricts the verifier input to fixed subsets of dimensions: 4-Dim keeps Neutrality, Correctness, Completeness, and Precision, while 2-Dim keeps Neutrality and Correctness. Label-only SFT is listed as a reference point; it uses a different prompt and emits no dimension scores.

Label-only SFT reaches 95.9% accuracy on Weibo21; introducing structured reasoning costs 13.0%, and the consistency term recovers 12.1% of that cost, so GSC retains label-only accuracy while making the emitted scores functional. We now ask whether the gain is carried by the dimensional score signal, by the gold label, or by the output format.

The Reasoning-only, Random-G and Shuffled-G block shows that the label and format alone cannot explain the full gain. Removing consistency supervision causes a 12.1% drop from GSC, and corrupting the verifier-side score signal still hurts even though the model sees the same gold label and

Variant	Acc	Macro-F1	Δ Acc	95% CI	McNemar p	Flips
Label-only SFT	0.9595	0.9595	-1.0%	$[-2.7\%, +0.7\%]$	0.314	27/36
Reasoning-only	0.8291	0.8274	+12.1%	$[+10.2\%, +15.4\%]$	2.37×10^{-21}	141/24
Random-G	0.9157	0.9157	+3.4%	$[+2.1\%, +6.2\%]$	1.11×10^{-4}	66/28
Shuffled-G	0.9025	0.9025	+4.7%	$[+3.3\%, +7.6\%]$	5.64×10^{-7}	75/25
Trainable-G	0.8981	0.8974	+5.2%	$[+3.1\%, +7.3\%]$	4.03×10^{-6}	75/28
4-Dim	0.8686	0.8675	+8.1%	$[+5.7\%, +10.6\%]$	1.33×10^{-10}	105/31
2-Dim	0.8565	0.8562	+9.3%	$[+7.7\%, +12.5\%]$	3.62×10^{-16}	114/22
GSC (ours)	0.9496	0.9496	—	—	—	—

Table 2: Paired mechanism analysis on Weibo21. Δ Acc, the confidence interval, and the McNemar test are all computed against full GSC; flips are control-wrong/GSC-right over GSC-wrong/control-right. The g_ϕ -side controls hold the backbone, prompt, output schema, training steps, and gold veracity label fixed. Label-only SFT uses a different prompt and emits no dimension scores.

emits the same structured schema. This matters because Random-G and Shuffled-G do not make the label token harder to learn; they only break the semantic relation between the current instance and the score vector checked by g_ϕ .

The Trainable-G and Reduced-Dim block shows that the verifier must remain a stable human-calibrated constraint. Jointly training g_ϕ reduces accuracy by 5.2%, while restricting the verifier to fewer score dimensions also weakens performance, with both 2-Dim and 4-Dim trailing the full six-

Variant	Acc	Macro-F1	Δ Acc
No f_ψ	0.7891	0.7792	+16.1%
Random-pos	0.5703	0.5435	+37.9%
Prev-token	0.5742	0.5157	+37.5%
First-6	0.5391	0.5103	+41.1%

Table 3: f_ψ -side score-head probe on Weibo21, compared with full GSC.

dimensional GSC model. Thus the effect depends on score semantics that are fixed, multidimensional, and checkable during SFT. We fix $\beta = 0.7$ for the main runs and controlled variants; the sensitivity sweep is reported in Appendix A.2.

5.4 Score Head f_ψ Readout Analysis

We next probe how the readout path itself shapes the gain, beyond what g_ϕ checks. After isolating the verifier side, we ask whether any hidden representation can serve as the input to g_ϕ . Table 3 keeps the same backbone, frozen verifier, prompt template, and $\beta = 0.7$, changing only the readout path. **No f_ψ** removes the learned adaptor and derives scalar scores from the LM head’s score-vocabulary probabilities, while **Random-pos**, **Prev-token**, and **First-6** keep f_ψ but read from positions that are not the score tokens.

Table 3 separates two failure modes. Removing f_ψ drops accuracy below *Reasoning-only*, showing that score-vocabulary logits are a poor substitute for a learnable continuous bridge. Reading from the wrong positions is even more damaging, and Random-pos, Prev-token, and First-6 all collapse far below the no-consistency baseline. The one-token-offset control is particularly diagnostic, since the model is still near the score field while the consistency signal fails almost as badly as random positions. The useful training signal is therefore carried specifically by the hidden states where the model commits to its score tokens.

5.5 Hidden-State Score Intervention

We further test whether score-position representations are behaviorally active during generation. For each model, we replace the six score-position hidden states with hidden states taken from opposite-label samples at a fixed layer, then continue generation and measure whether the final veracity label changes. This intervention preserves the prompt, output schema, and decoding procedure, while perturbing the residual stream at the positions where the model expresses the six scores before the final

Dataset	Model	Base Acc	L14 Flip	L20 Flip	L27 Flip
Weibo21	GSC	0.9496	17.27%	6.36%	0.00%
	Reasoning-only	0.8291	1.82%	0.91%	0.00%
Twitter	GSC	0.8973	6.00%	4.00%	0.00%
	Reasoning-only	0.6751	1.35%	0.00%	0.00%

Table 4: Hidden-state score intervention. Base Acc is normal generation accuracy; flip rate is the fraction of cases whose generated label changes after replacing the six score-position hidden states with opposite-label hidden states at the indicated layer. The Twitter run uses the same β .

SFT readout forms its prediction.

Table 4 shows that GSC is substantially more sensitive to score-position interventions than Reasoning-only. At layer 14, replacing the six score-position hidden states flips the generated label in 17.27% of GSC cases, compared with only 1.82% for Reasoning-only; the gap remains visible at layer 20 and disappears by layer 27. The middle-layer peak indicates that GSC uses score-position representations while integrating the structured signal; by the late layer the decision state has largely absorbed this information, so replacing the score positions alone no longer changes the already-formed label.

5.6 Score Discriminability

On the Weibo21 test set we compare dimension-wise scores across four training stages and report, for each dimension, the gap between the real and the fake group divided by the pooled within-group standard deviation, that is, Cohen’s d (Figure 4, Table 5). The consistency term raises this separation on all six dimensions, from 1.33 to 1.73 on average. GSC and Reasoning-only share the same backbone, prompt, output schema, training data and gold labels, and the consistency loss is the only difference between them. The same sequence speaks to the annotation step that relies on M_{grpo} . The GRPO alignment consumes 300 expert-annotated instances, a budget far below the later SFT stage, and reaches a separation of 0.73 on the test set; training on the 8,010 instances that it annotates raises the separation to 1.33. The machine-generated score annotations therefore carry a learnable signal, and the consistency term lifts the separation further to 1.73. Separation summarizes the between-group relation between scores and labels, so this section supplies evidence complementary to the mechanism reading drawn from the ablations, with per-dimension distributions in Appendix A.3.

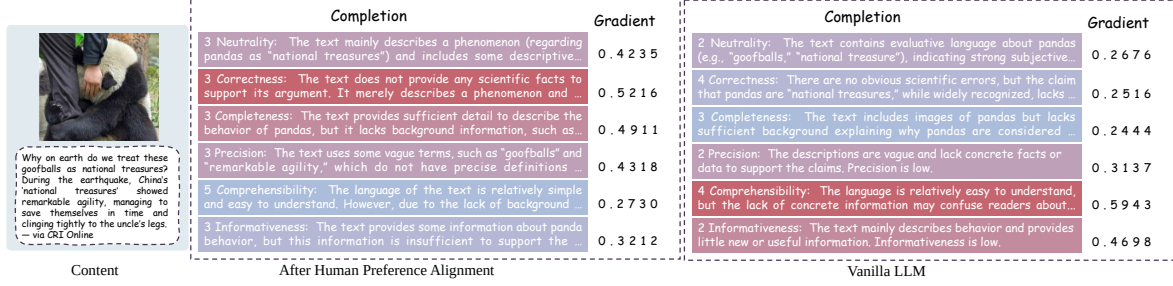


Figure 3: Gradient visualization of model-generated output paragraphs, where each paragraph corresponds to the analysis of a predicted text quality dimension. The score of each dimension is shown before the corresponding paragraph, and the gradient value denotes the average token-level gradient within the paragraph.

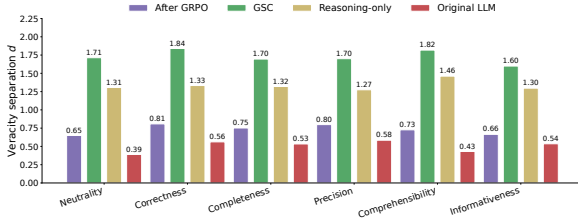


Figure 4: Per-dimension veracity separation on the Weibo21 test set, measured as Cohen’s d between the real and the fake group. Human expert score distributions appear in Figure 6.

Dimension	Mean score				Veracity separation d			
	GRPO	GSC	Reason.	Orig.	GRPO	GSC	Reason.	Orig.
Neutrality	4.66	4.65	4.59	4.13	0.65	1.71	1.31	0.39
Correctness	4.00	4.15	3.87	3.43	0.81	1.84	1.33	0.56
Completeness	4.15	4.26	3.98	3.64	0.75	1.70	1.32	0.53
Precision	3.89	4.09	3.78	3.43	0.80	1.70	1.27	0.58
Comprehensibility	5.71	5.39	5.19	4.82	0.73	1.82	1.46	0.43
Informativeness	4.17	4.06	3.82	3.50	0.66	1.60	1.30	0.54
Mean	4.43	4.44	4.21	3.82	0.73	1.73	1.33	0.51

Table 5: Per-dimension mean scores and veracity separation on the Weibo21 test set. The separation is Cohen’s d , the gap between the two veracity groups divided by the pooled within-group standard deviation. Human expert score distributions appear in Figure 6.

5.7 Qualitative Case Study

The qualitative pattern is consistent with the proposed mechanism, where alignment changes which reasoning dimensions the model relies on. Figure 3 visualizes gradient contributions from different dimension-specific reasoning segments to the predicted label token on a representative multimodal post. Gradients are computed with respect to the predicted label probability and averaged within each segment, so paragraph length does not dominate the visualization. The vanilla LLM places relatively more attribution on later prompt segments, especially *Comprehensibility* and *Informativeness*, which can be satisfied by fluent but shallow descriptions. After alignment, attribution shifts toward *Correctness*, a more direct veracity signal.

5.8 Summary of Mechanism Evidence

Together with the hidden-state intervention in Table 4, the qualitative case study suggests that GSC changes both where score information is represented and which dimension-specific reasoning segments the label prediction is sensitive to. Taken together, the experiments identify a specific mechanism behind the gain. The label and output format are controlled, yet Reasoning-only, Random-G, and Shuffled-G trail GSC under paired tests; making g_ϕ trainable weakens the result, showing that a frozen human-calibrated score-to-label relation matters; Table 3 shows that the verifier must be driven through a trainable score-token readout; Table 4 shows that intervening on score-position hidden states changes GSC generations far more often than Reasoning-only; and Table 5 shows that GSC widens the veracity separation carried by the emitted scores. We therefore interpret the gains as score-semantic calibration with behaviorally active score-position representations, with Twitter reproducing the same ordering among LLM baselines in Table 1, and Weibo21 providing the paired statistical evidence.

6 Conclusion

We have shown that structured LLM outputs in misinformation detection are not necessarily functional. Adding structured-reasoning targets to label supervision costs accuracy, because the format of the trace can be satisfied without binding the intermediate scores to the veracity decision. We propose GSC to close this gap with low supervision, by training a compact human-calibrated verifier on a small set of expert-annotated posts, freezing it, and using a trainable score head to deliver the verifier signal at the score-token hidden states the LLM commits to during generation. The consistency term recovers that cost and restores accuracy to a

level statistically indistinguishable from label-only supervision.

Across label- and format-matched controls, score-token positional probes, and hidden-state interventions, the gain is jointly attributable to the semantic validity of the frozen score-to-label prior and to the behavioral activity of the score-token representations, and the direction reproduces across datasets. No additional label term is introduced. The only new supervision is a consistency loss at the score positions, and the benefit surfaces on the label. Overall, we hope this work encourages the community to treat structured reasoning in LLM-based misinformation detection as a behavioral property of the model rather than as a surface feature of its outputs.

Limitations

This study focuses on a controlled setting for examining whether structured LLM reasoning can be made verifier-consistent through human-calibrated score consistency. Our experiments explore the 7B backbone in depth, and other model scales are not explored to the same depth; the conclusions we draw therefore concern this scale, and how the mechanism behaves for substantially smaller or larger models remains open. While the experiments cover representative multimodal fake-news benchmarks and a strong instruction-tuned VLM backbone, future work could extend the evaluation to broader datasets, languages, event domains, model families, and model scales. The human annotation budget and the six-dimensional rubric are intentionally compact to study a low-supervision regime; larger or more diverse annotation protocols may further clarify how verifier quality, rubric design, and domain shift affect performance. We hope the proposed consistency framework can support future work on more reliable, human-in-the-loop misinformation detection systems.

Ethical Considerations

This work studies misinformation detection on public benchmark datasets and uses human annotations only to calibrate information-quality scores for research purposes. Because these benchmarks are derived from social-media posts, they may contain user-generated content; we rely on the original datasets’ anonymization and release conditions, do not collect new social-media data, and use the data only for research. The proposed system is

intended as decision support for human reviewers; real-world moderation or fact-checking use should keep humans in the loop and account for dataset bias, cultural context, and potential harm from false positives.

Acknowledgments

This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2022D01D81, Grant No. 2025D01B159); the Tianshan Talent Training Program (Grant No. 2023TSYCCX0041); the Outstanding Member Program of the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Grant No. Y2023118); the Key Research and Development Program of Xinjiang Uygur Autonomous Region (Grant No. 2023B03024, Grant No. 2024B03026); the Key Science and Technology Innovation Program of the Chinese Academy of Sciences (Grant No. KGFZD-145-26-44); the Youth Talents Support Project of Xinjiang Uygur Autonomous Region (Grant No. 2023TSYCQNTJ0037); the “Tianshan Elite” Science and Technology Top-notch Youth Talents Program (Grant No. 2023TSYCCX0044); and the Xinjiang Talent Development Fund (Grant No. XJRC-2025-KJ-PY-KJLJ-048). Generative AI was used to assist with text editing and polishing; all ideas, analyses, and experimental designs were proposed, conducted, and verified by the authors. We thank the anonymous reviewers and the area chairs of ACL Rolling Review and EMNLP 2026 for their constructive feedback, which significantly helped improve this work.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. 2024. [Crowd-sourced fact-checking: Does it actually work?](#) *Information Processing & Management*, 61(5):103792.
- Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopolou, and Yiannis Kompatsiaris. 2018. [Detection and visualization of misleading content](#)

- on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.
- Davide Ceolin, Kevin Roitero, and Furong Guo. 2023. [Predicting crowd workers performance: An information quality case](#). In *Web Engineering - 23rd International Conference, ICWE 2023, Alicante, Spain, June 6-9, 2023, Proceedings*, volume 13893 of *Lecture Notes in Computer Science*, pages 75–90. Springer.
- Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. 2022. [Cross-modal ambiguity learning for multimodal fake news detection](#). In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2897–2905. ACM.
- Tsun-Hin Cheung and Kin-Man Lam. 2023. [Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking](#). In *Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023, Taipei, Taiwan, October 31 - Nov. 3, 2023*, pages 846–853. IEEE.
- Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. [The effects of crowd worker biases in fact-checking tasks](#). In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2114–2124. ACM.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron C. Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. [V-Star: Training verifiers for self-taught reasoners](#). In *First Conference on Language Modeling (COLM)*.
- Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. [Bad actor, good advisor: Exploring the role of large language models in fake news detection](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 22105–22113. AAAI Press.
- Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. [Synergizing LLMs with global label propagation for multimodal fake news detection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 1426–1440. Association for Computational Linguistics.
- International Organization for Standardization. 2008. ISO/IEC 25012:2008 software engineering — software product quality requirements and evaluation (SQuARE) — data quality model. Technical report, International Organization for Standardization, Geneva, Switzerland.
- J. M. Juran. 1992. *Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services*. The Free Press, New York.
- An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. [Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 18426–18434. AAAI Press.
- Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. [Misinformation and its correction: Continued influence and successful debiasing](#). *Psychological science in the public interest*, 13(3):106–131.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Eddy Maddalena, Davide Ceolin, and Stefano Mizzaro. 2018. [Multidimensional news quality: A comparison of crowdsourcing and nichesourcing](#). In *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018*, volume 2482 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. [MDFEND: multi-domain fake news detection](#). In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3343–3347. ACM.
- Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. [Let silence speak: Enhancing fake news detection with generated comments from large language models](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1732–1742. ACM.
- Dan Saattrup Nielsen and Ryan McConville. 2022. [Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset](#). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3141–3153. ACM.
- OpenAI. 2023. [Gpt-3.5 turbo](#).
- OpenAI. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- OpenAI. 2026. [OpenAI GPT-5 system card](#). *arXiv preprint arXiv:2601.03267*.
- Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. [Partisan polarization is the primary](#)

- psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015.
- Sungwon Park, Sungwon Han, Xing Xie, Jae-Gil Lee, and Meeyoung Cha. 2025. [Adversarial style augmentation via large language model for robust fake news detection](#). In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pages 4024–4033. ACM.
- Kevin Roitero, Dustin Wright, Michael Soprano, Isabelle Augenstein, and Stefano Mizzaro. 2025. [Efficiency and effectiveness of LLM-based summarization of evidence in crowdsourced fact-checking](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 457–467. ACM.
- Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. [Spotfake: A multi-modal framework for fake news detection](#). In *Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*, pages 39–47. IEEE.
- Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. [The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale](#). *Information Processing & Management*, 58(6):102710.
- Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. 2007. [A framework for information quality assessment](#). *Journal of the American Society for Information Science and Technology*, 58(12):1720–1733.
- Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. [DAPT: domain-aware prompt-tuning for multimodal fake news detection](#). In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 7902–7911. ACM.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Jacky Visser, John Lawrence, and Chris Reed. 2020. [Reason-checking fake news](#). *Communications of the ACM*, 63(11):38–40.
- Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. [DELL: generating reactions and explanations for llm-based misinformation detection](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2637–2667. Association for Computational Linguistics.
- Yair Wand and Richard Y. Wang. 1996. [Anchoring data quality dimensions in ontological foundations](#). *Communications of the ACM*, 39(11):86–95.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. [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, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.
- Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, and Xiaoyuan Jing. 2025. [Balanced multi-modal learning with hierarchical fusion for fake news detection](#). *Pattern Recognition*, 164:111485.
- Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. [Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3367–3378. ACM.
- Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian J. McAuley. 2024. [Small models are valuable plug-ins for large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 283–294. Association for Computational Linguistics.
- Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. [Bootstrapping multi-view representations for fake news detection](#). In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 5384–5392. AAAI Press.
- Mohammad Yousuf. 2025. [Mediating the truth: Influences of routines on legacy news media fact-checking](#). *Journalism Practice*, 19(1):93–113.
- Fei Yu, Anningzhe Gao, and Benyou Wang. 2024. [OVM, outcome-supervised value models for planning in mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 858–875. Association for Computational Linguistics.
- Yaozeng Zhang, Jing Ma, and Yuguang Jia. 2025. [MCAN: multimodal cross-aware network for fake](#)

news detection by extracting semantic-physical feature consistency. *The Journal of Supercomputing*, 81(1):299.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. **SAFE: similarity-aware multi-modal fake news detection**. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, volume 12085 of *Lecture Notes in Computer Science*, pages 354–367. Springer.

A Mechanism and Calibration Evidence

We order the appendix by proximity to the paper’s central claim. We first present the evidence that the consistency loss affects training behavior and generated score distributions, then the human score space that calibrates the verifier, followed by baselines, implementation details, and prompt templates.

A.1 Consistency Objective and Module Details

Consistency mask. Because $\mathcal{D}_{\text{LLM}}$ contains a small fraction of fallback instances whose target sequences do not carry the six dimension scores produced by M_{grpo} (§C.1), the consistency term $\mathcal{L}_{\text{GSC}}$ is undefined on those instances, since no score-token positions exist in the target sequence at which the score head f_ψ can compute a six-dimensional prediction, and forcing g_ϕ to verify a zero-padded or imputed score vector would inject noise into a signal that the rest of the design is built to keep clean. We therefore gate the consistency term with the annotation-time binary mask m_x :

$$\mathcal{L}_{\text{total}}(x) = \mathcal{L}_{\text{LM}}(x) + m_x \beta \mathcal{L}_{\text{GSC}}(x). \quad (6)$$

Strict-success instances therefore train both terms with the weight $\beta = 0.7$ fixed throughout this paper; fallback instances train only the language-modeling term, contributing to label-token fluency without participating in the verifier-mediated consistency objective.

The forward-pass implementation is straightforward, where every training batch reads m_x once per instance from the corpus metadata produced during score annotation and then broadcasts it to mask out the $\mathcal{L}_{\text{GSC}}$ contribution of fallback instances when reducing the per-instance loss to the batch mean. We do not anneal the mask and we do not re-evaluate the strict-success criterion at training time; m_x is fixed at annotation time and stays fixed throughout final SFT.

Verifier g_ϕ pre-training. The verifier is a one-hidden-layer MLP that maps a six-dimensional score vector to two-way veracity logits. Its computation graph is $\text{Linear}(6 \rightarrow 64) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.0) \rightarrow \text{LayerNorm}(64) \rightarrow \text{Linear}(64 \rightarrow 2)$, with no input LayerNorm and a single post-activation LayerNorm before the output layer, for a total of 706 trainable parameters. We train g_ϕ from scratch on $\mathcal{D}_{\text{human}}$ (600 expert-annotated instances with per-dimension score vectors and gold veracity labels; see Appendix B) using AdamW with learning rate 3×10^{-4} , weight decay 1×10^{-4} , batch size 64, and seed 44, for up to 400 epochs with validation every 5 epochs and no class-weighted loss. We retain the model state that maximizes validation weighted-F1, breaking ties by validation accuracy and then by lower validation loss; on $\mathcal{D}_{\text{human}}$ this selection lands at epoch 5 (validation accuracy 72.69%, weighted-F1 71.85%, macro-F1 71.19%). After selection g_ϕ is frozen for the remainder of the pipeline and never receives a gradient during LLM SFT. This freezing is the load-bearing design choice analyzed in §4.2.

Score head f_ψ . The score head reads the LLM’s final readout hidden states at the six score-token positions and projects each dimension into a scalar score with an *independent* per-dimension sub-head; we do not share parameters across dimensions. Concretely, f_ψ is a stack of six sub-heads $f_\psi^{(1)}, \dots, f_\psi^{(6)}$, where each $f_\psi^{(k)} : \mathbb{R}^{3584} \rightarrow \mathbb{R}$ has the computation graph $\text{LayerNorm}(3584) \rightarrow \text{Linear}(3584 \rightarrow 256) \rightarrow \text{GELU} \rightarrow \text{Dropout}(0.1) \rightarrow \text{Linear}(256 \rightarrow 1)$, with 3584 being the hidden size of Qwen2.5-VL-7B (Bai et al., 2025). The k -th sub-head consumes only the hidden state at the k -th score-token position and emits a scalar $\hat{s}_k$; the six outputs are concatenated to form $\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_6) \in \mathbb{R}^6$, which is then read by the frozen verifier g_ϕ to compute $\mathcal{L}_{\text{GSC}}$. The six sub-heads together hold $\approx 5.55\text{M}$ trainable parameters.

Unlike g_ϕ , f_ψ is *not* pre-trained on $\mathcal{D}_{\text{human}}$: its parameters are randomly initialized and trained *only* during final SFT, jointly with the LoRA adapters under $\mathcal{L}_{\text{LM}} + \beta \cdot \mathcal{L}_{\text{GSC}}$ as gated by Eq. 6; f_ψ receives no gradient during GRPO alignment, label-free annotation, or g_ϕ pre-training. Its optimizer, learning rate, scheduler, and warmup are inherited from the final SFT configuration in Table 10; no separate f_ψ -specific schedule is used. In all main runs no direct score-regression term and

β	Acc	Real F1	Fake F1
0.15	0.871	0.874	0.867
0.20	0.890	0.889	0.892
0.30	0.891	0.891	0.891
0.50	0.909	0.914	0.905
0.60	0.914	0.917	0.910
0.70	0.929	0.933	0.925
0.80	0.923	0.927	0.918
1.00	0.902	0.904	0.899

Table 6: β sensitivity on a Weibo21 validation set of $n = 1,159$ instances drawn from the training split, with all expert-annotated instances excluded. The argmax stays at $\beta = 0.7$.

no label-token reweighting are applied, so that the only auxiliary term added to $\mathcal{L}_{\text{LM}}$ is $\beta \cdot \mathcal{L}_{\text{GSC}}$ as written in Eq. 6.

A.2 β Sensitivity

Table 6 reports a β sweep on Weibo21 with $\beta \in \{0.15, 0.20, 0.30, 0.50, 0.60, 0.70, 0.80, 1.00\}$. The sweep uses the same hyperparameter recipe as Table 10. At $\beta = 0$ the consistency term vanishes and the objective collapses to Reasoning-only, so the grid starts above zero. The sweep runs on a validation set of $n = 1,159$ instances drawn from the Weibo21 training split, which excludes all expert-annotated instances and was held out of the GSC runs reported here.

Accuracy rises from 87.1% at $\beta = 0.15$ to a peak of 92.9% at $\beta = 0.70$ and falls to 90.2% at $\beta = 1.00$. The region $\beta \in [0.5, 0.8]$ stays within two points of the peak, so the preferred range is broad rather than narrow. We adopt $\beta = 0.7$ for all reported runs.

A.3 Training-Stage Score Distributions

The per-dimension histograms expand the separation figures of §5.6 to the level of distributions. GSC and Reasoning-only use the same training data and the same output schema, and GSC spreads its samples over more score levels, with the most frequent level holding 33.3% of the samples on average against 38.4% for Reasoning-only, on all six dimensions. The GRPO stage and the original LLM place 23.5% and 25.5% at their most frequent level, with separations of 0.73 and 0.51. This agrees with the reading that the consistency term acts on the generated scores and on the final label token together.

Variant	3B	7B	72B
Reasoning-only	0.831	0.829	0.841
Trainable-G	0.869	0.898	0.863
GSC	0.930	0.950	0.958
GSC – Reasoning-only	+9.9%	+12.1%	+11.7%

Table 7: Weibo21 accuracy across backbone scales, with $\beta = 0.7$ transferred without retuning. The cost of removing the consistency term persists at every scale.

A.4 Scale Effect

The controlled comparison in Table 2 uses a 7B backbone. To check whether the accuracy cost of structured reasoning is a property of that scale, we repeat the Reasoning-only and GSC settings on Qwen2.5-VL-3B and Qwen2.5-VL-72B under the same recipe, with $\beta = 0.7$ carried over unchanged, a fixed epoch budget, and the final checkpoint taken without selection. Table 7 reports Weibo21 accuracy.

Two patterns hold at every scale. First, removing the consistency term costs accuracy at all three sizes, and the gap does not shrink as the backbone grows: Reasoning-only trails GSC by 9.9% at 3B, 12.1% at 7B, and 11.7% at 72B. Structured reasoning supervision on its own is therefore not a small-model artifact. Second, Trainable-G stays below GSC at all three sizes, so freezing the verifier remains necessary even at 72B. What scale improves is the overall discrimination ability of the backbone; the coupling between the emitted scores and the label does not appear on its own, and in every setting we tested it is the consistency term that establishes it.

B Human Annotation and Score Space

This appendix expands on the human-annotation protocol summarized in §3 and documents the score space used to calibrate the frozen verifier. We describe the six information quality dimensions, the sampling protocol used to construct $\mathcal{D}_{\text{human}}$, the rating scale and aggregation rule, the inter-annotator agreement, and auxiliary score-distribution evidence from human, GPT-5.2, and Qwen2.5-VL-7B scores.

B.1 Annotation Protocol and Dimension Definitions

Dimension definitions. We adopt six of the seven dimensions of Soprano et al. (2021), grounded in the ISO 25012 information-quality

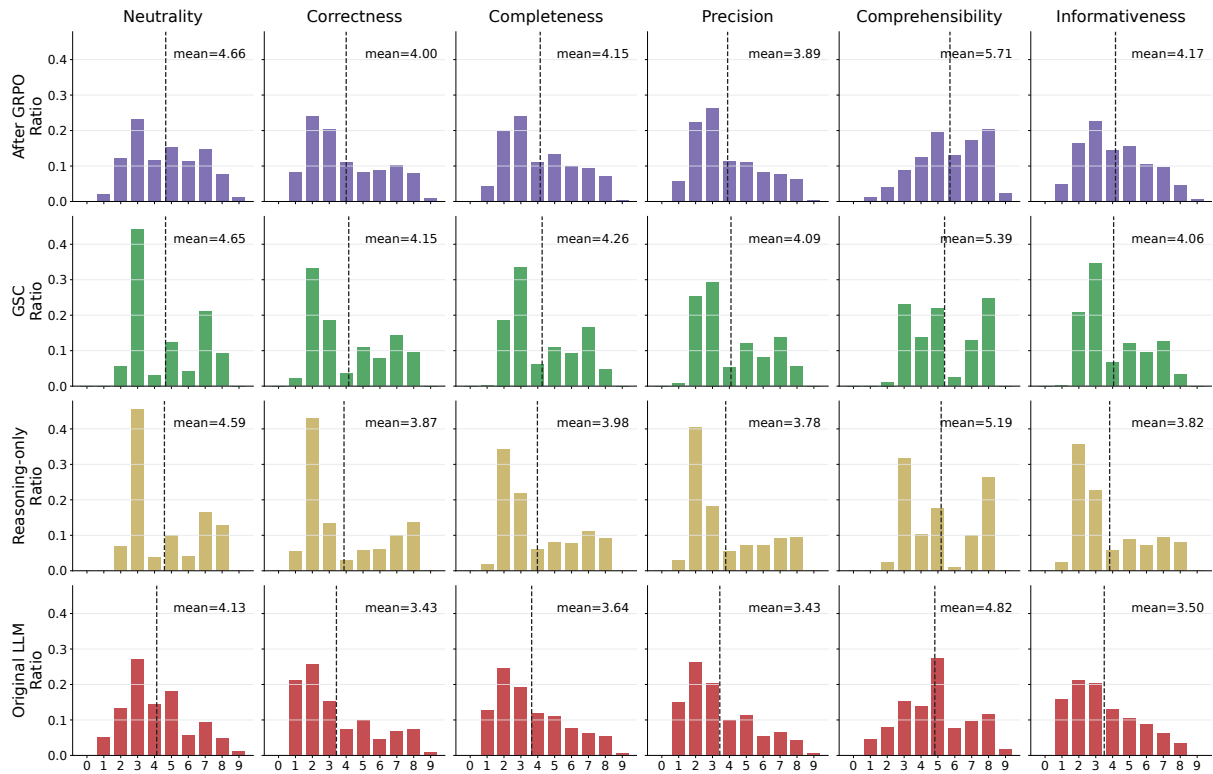


Figure 5: Weibo21 test-set score distributions across four training stages. Human expert score distributions appear in Figure 6.

model (International Organization for Standardization, 2008) and validated for credibility judgments by both experts and crowd workers (Maddalena et al., 2018; Barbera et al., 2024). We omit Speaker’s Trustworthiness because social-media news posts are frequently reposted or redistributed, making it cumbersome to trace the original publisher from the post alone. The definitions used by our annotators are:

- **Neutrality.** Whether a post adopts an objective, factual tone or primarily expresses subjective viewpoints, emotion, or stance. Emotive hashtags (e.g., #scary), exclamatory punctuation, and overt opinion markers (“I think,” “obviously”) push the score down.
- **Correctness.** Whether the content is consistent with commonly accepted real-world constraints and is internally self-consistent, judged *without* external fact-checking. This is a plausibility-style judgment of the text as read, not a verdict on the underlying event.
- **Completeness.** Whether the essential elements for understanding the event (actor, time, location, cause, outcome) are present. Hash-tags do not count as elements; an image may

complete the description if the text explicitly refers to it.

- **Precision.** The degree to which claims are articulated with specific temporal, spatial, numerical, or named details. Note that a fabricated post can be highly precise; precision and truth are not coupled by definition.
- **Comprehensibility.** How easily the post can be read and understood as a linguistic artifact, independent of whether its content is true or complete. Foreign-language posts are not penalized for being in a foreign language; only intrinsic ambiguity, broken syntax, or unclear references reduce the score.
- **Informativeness.** Whether the post conveys substantive, verifiable content beyond exclamations, hashtags, URLs, or pure opinion. Hashtags, image-pointer phrases (“see this!”), and bare URLs do not contribute.

The annotator-facing rubric additionally provides anchor descriptions for each integer point on the 1-to-9 scale, worked examples, and a self-check list.

Sampling and annotation. $\mathcal{D}_{\text{human}}$ consists of 600 posts sampled from the *training* splits of Weibo-21 (Nan et al., 2021) and Twitter (Boididou et al., 2018). Sampling is uniform-random within each platform and is stratified by veracity to obtain a balanced subset:

- Weibo-21: 150 real + 150 fake = 300 posts.
- Twitter: 150 real + 150 fake = 300 posts.

We impose no topic or time restrictions, and we verify that no sampled training post coincides with any post in the evaluation splits used in §5. The resulting subset is disjoint from the Weibo21 and Twitter evaluation splits used for the main report.

Five annotators participated; all are graduate-level researchers familiar with social-media misinformation. Each annotator receives the same rubric, a short calibration session on 20 pilot posts (not part of $\mathcal{D}_{\text{human}}$), and rates each post on every dimension using a 1-to-9 integer scale. The veracity label is also recorded for each post but is collected independently of the dimension scores and is not used as a target for any dimension. Every post in $\mathcal{D}_{\text{human}}$ is rated by all five annotators, and the post-level final score on each dimension is the per-annotator mean, which is the value supplied to the verifier g_ϕ during training (§B.1).

Rating scale and aggregation. We use a 1-to-9 integer scale on every dimension, with 5 as an explicit “cannot tell / default” value and odd integers (1, 3, 5, 7, 9) anchored to qualitative levels (worst, low, mid, high, best). Even integers are reserved for in-between calls. The rubric explicitly instructs annotators to avoid using 4 or 6 as fallbacks when uncertain, since uncertain cases default to 5.

For each post and dimension, the final score released to downstream pipelines is the per-dimension mean across the participating annotators, rounded to the nearest integer in $\{1, \dots, 9\}$. This rounded mean is the value supplied to g_ϕ during its pre-training (§4.2); the score head f_ψ is randomly initialized and not pre-trained on these targets (Appendix A.1).

B.2 Agreement and Score Distribution Evidence

We compute agreement on the full annotated subset for each dataset, namely 300 Twitter posts and 300 Weibo-21 posts, the entire $\mathcal{D}_{\text{human}}$ sample. Each post in both subsets is rated by every participating annotator on every dimension. Per-dataset

metrics are then averaged across the two datasets (arithmetic mean of the per-dataset Krippendorff α , Fleiss κ , score mean, score standard deviation, and pairwise disagreement rate).

For each dimension we report Krippendorff α , which is computed at the interval level appropriate for ordinal 1-to-9 ratings, and Fleiss κ , which treats each score as a separate category and is therefore conservative because it ignores the ordinal structure; we also report the score mean, standard deviation, and the fraction of annotator pairs that disagree by at least one scale point (*disagreement*).

Dimension	α	κ	mean	std	disagree
Neutrality	0.763	0.271	5.27	2.59	0.252
Correctness	0.668	0.232	5.98	1.87	0.296
Completeness	0.652	0.234	5.39	2.11	0.347
Precision	0.718	0.286	5.10	2.37	0.289
Comprehensibility	0.580	0.196	6.92	1.88	0.246
Informativeness	0.506	0.213	5.12	2.10	0.423

Table 8: Inter-annotator agreement on $\mathcal{D}_{\text{human}}$, per dimension. α is Krippendorff’s α at the interval level; κ is Fleiss’ κ on the raw 1-to-9 categories; disagreement is the fraction of annotator pairs differing by ≥ 1 scale point. Values are averaged across the Twitter ($n=300$) and Weibo-21 ($n=300$) subsets as described in §B.2.

The pattern in Table 8 is consistent with what has been previously reported for fine-grained truthfulness dimensions on social-media content (Barbera et al., 2024; Maddalena et al., 2018), where ordinal agreement (α) is moderate to substantial (0.51–0.76) while the raw categorical κ is markedly lower because a 9-category scale penalizes minor ordinal drift heavily. The lowest-ordinal-agreement dimensions are Comprehensibility ($\alpha = 0.58$) and Informativeness ($\alpha = 0.51$); these are also the two dimensions on which §3 reports the smallest distributional separation between real and fake posts, indicating that low IAA reflects intrinsic ambiguity of the construct on short, fragmentary posts. Conversely, Neutrality and Precision show the highest ordinal agreement ($\alpha = 0.76$ and 0.72), matching their cleaner separability in the score distributions.

We interpret these IAA numbers as evidence that the six dimensions, as defined in §B.1 and instantiated by the rubric, support reliable ordinal judgments in the moderate-to-substantial range required to train the small, frozen verifier g_ϕ , with each dimension contributing partial signal that the verifier combines. This is the empirical basis for our framing of the six dimensions as a *structural carrier* of LLM reasoning, restated from §3.

Score distributions. Figures 6–8 report per-dimension score distributions from human experts, GPT-5.2 (OpenAI, 2026), and Qwen2.5-VL-7B (Bai et al., 2025) on the Weibo-21 (Nan et al., 2021) and Twitter (Boididou et al., 2018) training subsets. The distributions support the main-text claim that the dimensions are useful structural carriers whose combination the verifier learns.

Human expert distributions. Human scores show clearer real–fake separation on Neutrality, Correctness, Completeness, and Precision, while Comprehensibility and Informativeness largely overlap across labels.

GPT-5.2 score distributions. GPT-5.2 preserves the broad direction on Neutrality and Correctness but compresses scores toward the middle of the scale, especially on the other four dimensions.

Qwen2.5-VL-7B score distributions. Before consistency training, Qwen2.5-VL-7B produces much flatter and less label-separating score distributions, motivating the need to bind generated scores to the frozen verifier.

Cross-annotator patterns. Across the three sources, Neutrality and Correctness are the most stable dimensions, Completeness and Precision are useful but more calibration-sensitive, and Comprehensibility/Informativeness are the least label-separating. This pattern matches the IAA results in Appendix B.2.

C Preference Alignment and Score Annotation

This appendix records how the small human-calibrated scoring set is expanded into the structured supervision corpus used for final SFT.

C.1 GRPO Preference Alignment and Label-Free Annotation

GRPO preference alignment. GRPO aligns the backbone LLM with the per-dimension scoring preferences encoded in $\mathcal{D}_{\text{human}}$, producing the preference-aligned policy M_{grpo} used for large-scale score annotation. We adopt Group Relative Policy Optimization (GRPO) as the preference-alignment algorithm; the only methodological constraint imposed by GSC is that the reward signal is computed *entirely* on the 600 human-annotated instances in $\mathcal{D}_{\text{human}}$, so that M_{grpo} ’s sense of which score profiles accompany real versus fake content is

grounded in expert annotation. The complete loss formulation $\mathcal{L}_{\text{GRPO}}$, the group-relative advantage, the PPO clipped ratio, and the three reward components (r_{format} , r_{answer} , r_{pref}) follow the standard GRPO recipe and are not repeated here; Table 10 collects the GRPO and final SFT configurations. The one component specific to this work is r_{pref} : within each group of K samples drawn for the same post, we rank the samples by the squared deviation of their six scores from the expert scores and award a decreasing bonus by rank. r_{pref} is therefore a *group-relative ranking* reward over score profiles, so this stage aligns M_{grpo} with expert *preferences* among score profiles while leaving the absolute scale of the expert scores free.

We deliberately do not re-tune the GRPO recipe, and treat this step as a black-box preference alignment whose output is consumed by score annotation, and any hyperparameter difference relative to the standard GRPO recipe would make downstream gains attributable to the alignment recipe as much as to GSC. The only design choice that this paper exposes at the alignment interface is the requirement that the reward source be restricted to $\mathcal{D}_{\text{human}}$; that constraint is what later allows the label-free annotation protocol to produce trajectories whose six-dimensional score profiles are themselves behaviorally meaningful, carrying information beyond agreement with a gold label.

Large-scale score annotation. The annotation step uses the preference-aligned M_{grpo} from §C.1 to construct the supervision corpus $\mathcal{D}_{\text{LLM}}$ on which the final model is fine-tuned. The gold label is withheld from the prompt during sampling, so M_{grpo} must produce a full structured response (reasoning trace, six dimension scores with per-dimension justifications, and a predicted veracity label) on the basis of the input post alone, with no shortcut to the correct answer.

Sampling protocol. For each training instance $x \in \mathcal{D}_{\text{train}}$, we run an ordered temperature schedule $T \in \{0.2, 0.5, 0.8, 1.0, 1.2\}$, drawing five samples at each temperature with the remaining decoding parameters held fixed, for a maximum of 25 candidate responses per instance. Decoding uses nucleus sampling with $\text{top-}p = 0.9$ and $\text{top-}k = 50$, and a limit of 1024 new tokens; the prompt template is the Structured Scoring Prompt (Appendix E.2) with the gold label field removed at sampling time. The temperature schedule is monotone non-decreasing so that earlier candidates cor-

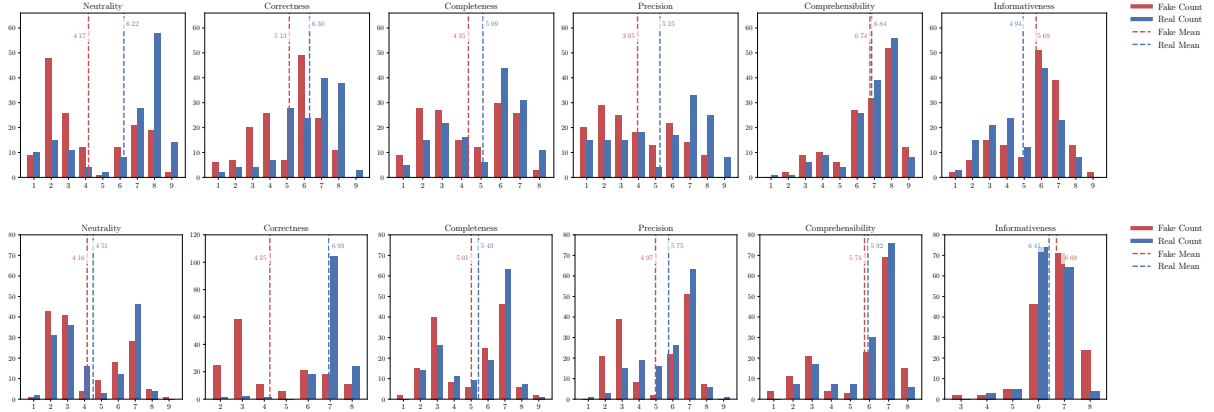


Figure 6: Human-expert score distributions on the Weibo-21 (top) and Twitter (bottom) training subsets of $\mathcal{D}_{\text{human}}$. Each row is a 1×6 panel, one panel per dimension; bars within a panel are stacked by veracity label.

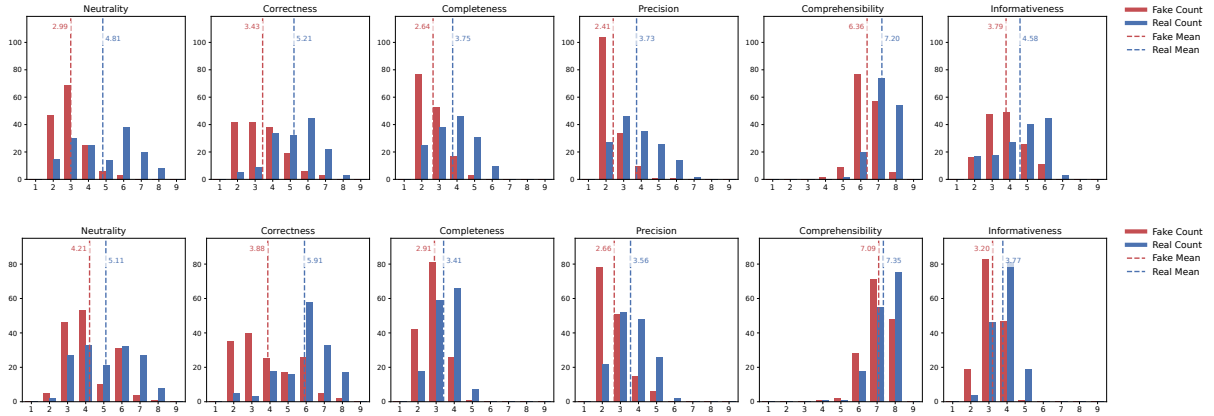


Figure 7: GPT-5.2 (OpenAI, 2026) score distributions on the Weibo-21 (top) and Twitter (bottom) training subsets. The same prompt template (Appendix B.1 rubric, rendered as a single-pass prompt) is used on both platforms.

respond to lower-entropy, higher-confidence roll-outs; when an early candidate already succeeds, the expensive high-temperature attempts are not consumed.

Strict-success criterion. A candidate is accepted as a *strict success* if and only if two conditions hold jointly. First, the predicted veracity label parsed from the candidate must match the gold label of the input. Second, all six score fields and their per-dimension justification fields must be present and parse cleanly under the JSON-like schema defined by the prompt template; partial parses, such as cases where only five of six scores are recoverable, are not accepted. For each instance we walk the 25 candidates in temperature and sample-index order and retain the *first* candidate that satisfies both conditions; the remaining candidates are discarded.

Fallback rule. If all 25 candidates fail the strict-success check, the instance falls back to a label-only target, where the target sequence is

constructed using the gold veracity label without any six-dimensional score field, and its consistency mask is set to $m_x = 0$. Strict-success instances carry the full target $\tilde{o} = (\text{reason}, s_1, \dots, s_6, \text{veracity})$ inherited from the accepted candidate, with the veracity token replaced by the gold label, and have $m_x = 1$.

Resulting corpus composition. Under this protocol the resulting $\mathcal{D}_{\text{LLM}}$ is dominated by strict-success instances. On Weibo21 approximately 97.5% of training instances yield a strict-success candidate within 25 samples and the remaining $\approx 2.5\%$ are fallback; on Twitter the corresponding ratio is approximately 95.8% strict-success and $\approx 4.2\%$ fallback. $\mathcal{D}_{\text{LLM}}$ contains 8,010 instances on Weibo21 and 13,366 on Twitter; on Twitter this equals the size of $\mathcal{D}_{\text{train}}$, while on Weibo21 it covers all but a small remainder of the training split documented in §5.1.

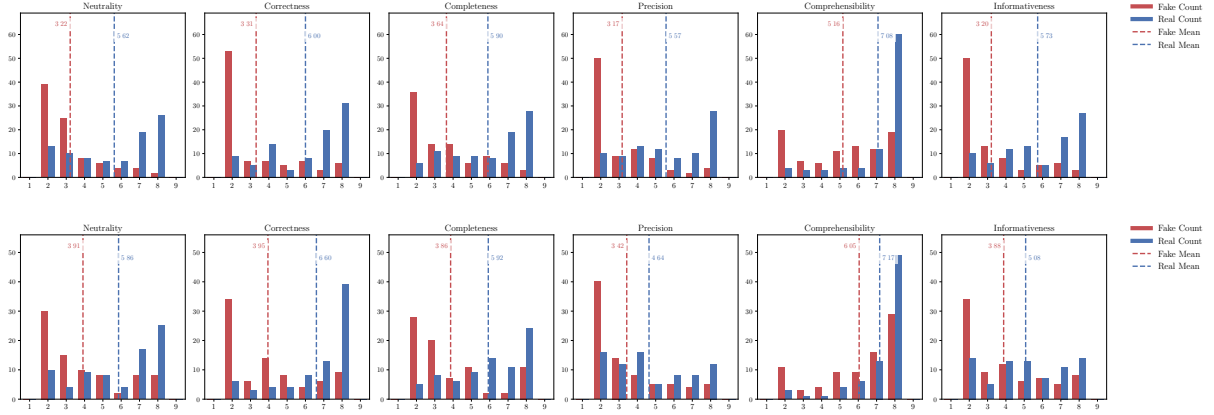


Figure 8: Qwen2.5-VL-7B (Bai et al., 2025) score distributions on the Weibo-21 (top) and Twitter (bottom) training subsets, obtained from the zero-shot backbone before any consistency training is applied.

Dimension	Human	Rank	Dev	R−H	D−H
Neutrality	4.11	4.69	4.55	0.58	0.45
Correctness	5.35	4.65	4.47	0.70	0.88
Completeness	4.91	4.73	4.16	0.18	0.75
Precision	4.35	4.39	3.84	0.05	0.51
Comprehensibility	6.76	6.11	6.47	0.65	0.29
Informativeness	4.41	4.78	4.10	0.37	0.30
Mean	—	—	—	0.42	0.53

Table 9: Mean dimension scores on the Twitter subset of the expert-annotated posts. Human is the average of the expert ratings; Rank and Dev are the scores emitted by generators aligned with the ranking and the deviation form of r_{pref} ; the last two columns are their absolute differences from Human.

Why label-free annotation. Revealing the gold label at sampling time would let M_{grpo} construct any six-dimensional score profile that is locally consistent with the prompted answer; under such a protocol the construction itself would couple the score fields to the final label, leaving the underlying reasoning unconstrained. Withholding the label forces the model to emit score profiles that are independently predictive of the truth, which is precisely the property the frozen verifier g_ϕ later certifies in the consistency term.

C.2 Reward Form for Preference Alignment

We further compare the ranking form of r_{pref} against a direct deviation reward on the expert scores. As shown in Table 9, the ranking form is more discriminative on the dimensions that bear most directly on veracity, while the deviation form falls slightly behind on those same dimensions. We therefore adopt the ranking form for r_{pref} .

D Implementation and Reproducibility Details

All variants reported in this paper fine-tune Qwen2.5-VL-7B (Bai et al., 2025) with LoRA. We deliberately fix the SFT budget across GSC, the controlled structured-output variants, and the label-only baseline so that differences in test accuracy isolate the score-supervision design under an equal optimization budget. The SFT variants therefore differ only in the source of the score target used by the consistency term. The label-only baseline omits scores entirely and uses a label-only prompt template; all structured variants share the same six-dimensional rubric prompt described in Appendix E.2. Table 10 lists both the GRPO alignment and final SFT configurations.

E Baselines and Prompt Templates

We place baseline descriptions and prompt templates last because they serve as experimental reference material for reproducing the setup.

E.1 Baseline Details

We group the baselines in Table 1 into multimodal detectors and LLM-based methods.

Multimodal detectors. EANN (Wang et al., 2018) uses event-category classification as an auxiliary task for fake news detection. SAFE (Zhou et al., 2020) extracts textual features with TextCNN and incorporates cross-modal similarity for veracity prediction. SpotFake (Singhal et al., 2019) combines BERT-based textual features with VGG-based visual features. CAFE (Chen et al., 2022) introduces ambiguity-aware multimodal fusion to improve robustness. BMR (Ying et al., 2023) dynam-

GRPO preference alignment	
Component	Setting
Backbone	Qwen2.5-VL-7B (Bai et al., 2025)
Adapter	LoRA on all linear layers; rank 256; $\alpha = 256$
Optimizer	AdamW
Learning rate	5×10^{-5}
LR scheduler	constant
Training budget	225 GRPO steps
Batching	group size $K = 8$; gradient accumulation 16 (Weibo21) / 32 (Twitter)
Regularization	KL coefficient $\beta_{KL} = 0.1$; PPO clip $\epsilon = 0.2^\dagger$
Sampling	rollout temperature 0.9
Reward source	$\mathcal{D}_{\text{human}}$ only
Reward weights	$r_{\text{format}} = 1$; $r_{\text{answer}} = 10$; $r_{\text{pref}} = (8, 4, 2, 1, 0.5, 0.25, 0.125, 0.05)$
Mixed precision	bfloat16
Hardware	$2 \times \text{A100}$ (80GB)

Final SFT	
Component	Setting
Backbone	Qwen2.5-VL-7B (Bai et al., 2025)
Adapter	LoRA; rank 64; $\alpha = 128$; dropout 0.05
Optimizer	AdamW
Learning rate	1×10^{-4}
LR scheduler / warmup	constant / 0.0
Training budget	3 epochs
Batching	per-device batch size 1; gradient accumulation 16
Regularization	max gradient norm 1.0; activation recomputation disabled
Input processing	max pixels 50176
Prompt templates	Structured Scoring Prompt / Label-Only Classification Prompt
Seeds	seed 42; data seed 42
Mixed precision	bfloat16
Hardware	single A100 (80GB)

Table 10: Training configurations for GRPO preference alignment and final SFT. † marks default values from TRL 0.27.0 and the Qwen2.5-VL reference implementation.

ically re-weights and refines multimodal features. **MCAN** (Zhang et al., 2025) combines semantic features, writing-style cues, and image-manipulation cues through cross-attention based multimodal fusion. **BMLHF** (Wu et al., 2025) addresses modality imbalance with hierarchical intra- and inter-modal fusion. **DAPT** (Tong et al., 2025) performs domain-aware prompt tuning with domain debiasing and uncertainty-aware multimodal fusion.

LLM-based methods. **GPT-3.5-turbo** (OpenAI, 2023), **GPT-4o** (OpenAI, 2024), and **GPT-5.2** (OpenAI, 2026) are evaluated in a zero-shot setting with the same task instruction format. **SuperICL** (Xu et al., 2024) incorporates auxiliary predictions from a smaller model into in-context learning. **ARG** (Hu et al., 2024) augments classifier predictions with LLM-generated reasoning. **Qwen2.5-VL-7B (zero-shot)** evaluates the backbone without task-specific fine-tuning. **Qwen2.5-VL-7B (SFT, label-only)** uses the same backbone

and final veracity supervision as GSC but omits the six score outputs and the consistency loss; its accuracy is reported in Table 2 as the reference point for the paired analysis.

E.2 Prompt Templates

This appendix reproduces the two prompt templates used in the main and ablation experiments, as referenced from Appendix C. Figure 9 contains the *Structured Scoring Prompt*, shared by GSC and the controlled structured-output variants. Figure 10 contains the *Label-Only Classification Prompt*, used by the Qwen2.5-VL-7B SFT label-only baseline reported in Table 2. All other hyperparameters are held fixed across the two templates (see Appendix D); only the structured output specification differs.

Prompt 1: Structured Scoring Prompt

System Prompt: You are an assistant for evaluating content veracity and objectivity, responsible for systematically analyzing the given content and assessing its truthfulness and credibility. You must strictly distinguish between factual statements and subjective opinions or emotionally driven expressions.

Please focus on the following six evaluation dimensions. First, carefully read the textual content. If an image is provided (i.e., not marked as “NO IMG”), analyze the image jointly with the text; otherwise, analyze the text only.

Conduct step-by-step reasoning and evaluate the credibility of the content along the six dimensions specified below.

The output format is as follows: <think> Please analyze the credibility of the content step by step, reasoning in detail along the following dimensions:

1. **Neutrality** - Determine whether the text primarily presents verifiable facts or expresses opinions, ideological stances, emotions, or speculation. - If the content is strongly fact-based and objectively descriptive → high neutrality (7–9); if it contains a mixture of factual statements and subjective elements → medium (4–6); if it is dominated by opinions or emotional expressions → low neutrality (1–3). - Note: Do not judge neutrality solely based on tone; assess whether the claims can be externally verified.

2. **Correctness** - Examine the consistency between the content and established scientific knowledge, domain-specific facts (e.g., natural or social sciences, historical records), and official or authoritative sources. - If the content clearly aligns with common knowledge and scientific principles → high (7–9); if it contradicts known facts or contains explicit errors → low (1–3); if evidence is insufficient or the statements are largely neutral → medium (4–6).

3. **Completeness** - Assess whether causal relationships, temporal sequences, and key details are clearly presented and internally consistent, and whether essential contextual information is missing. - If the text provides sufficient details, a coherent logical chain, and consistency with images and known facts → high (7–9); if there are logical gaps or missing critical information → low (1–3).

4. **Precision** - Evaluate whether the statements are specific and concrete (e.g., including numbers, dates, sources, or proportions), rather than vague or generalized. - If wording is precise and details are clearly specified → high (7–9); if the content is vague, exaggerated, overly general, or lacks concrete evidence → low (1–3).

5. **Comprehensibility** - Judge whether the content is easy to understand, well-structured, and clearly expressed, allowing readers to grasp the intended meaning without difficulty. - If the language is concise, logically organized, and free of significant ambiguity → high (7–9); if the text is verbose, obscure, grammatically flawed, or difficult to follow → low (1–3); otherwise assign a medium score (4–6).

6. **Informativeness** - Assess whether the content provides new, useful information rather than merely repeating common knowledge, clichés, or empty conclusions. - If the text contains concrete information that supports inference or further analysis → high (7–9); if it mainly consists of uninformative statements or widely known facts → low (1–3); otherwise assign a medium score (4–6).

</think>

Neutrality Score: <neutrality>[1–9]</neutrality>

Scientific Factual Correctness Score: <correctness>[1–9]</correctness>

Completeness Score: <completeness>[1–9]</completeness>

Precision Score: <precision>[1–9]</precision>

Comprehensibility Score: <comprehensibility>[1–9]</comprehensibility>

Informativeness Score: <informativeness>[1–9]</informativeness>

Content Veracity: <veracity>[True / False]</veracity>

User Prompt: Please assess the veracity and credibility of the following content. Analyze the text (and image, if provided) step by step, strictly distinguishing factual statements from subjective opinions. Evaluate the content along six dimensions—neutrality, factual correctness, completeness, precision, comprehensibility, and informativeness—and assign a score from 1 to 9 for each dimension. Finally, determine whether the content is true or false.

Context: News Text: [the given news text], Image: [the given news image]

Figure 9: Structured Scoring Prompt used by GSC and all controlled variants. The six dimension scores appear at fixed positions in the target sequence and are consumed by the verifier described in Section 4.

Prompt 2: Label-Only Classification Prompt

System Prompt: You are an assistant for assessing content veracity and objectivity. Your task is to conduct a systematic analysis of the given content and determine its factual accuracy and overall credibility. You must strictly distinguish between factual statements and subjective opinions or emotional expressions.

Your assessment should focus on identifying verifiable claims and evaluating their credibility. Please carefully read the provided text first. If an accompanying image is provided (i.e., not marked as "NO IMG"), incorporate visual evidence into your analysis. If no image is provided, base your assessment solely on the textual content.

Conduct your reasoning step by step, and analyze the credibility of the content in detail according to the following criteria.

Output format:

<think>

1. Identification of key factual claims (list each claim separately)
2. Verifiability analysis for each factual claim
3. Detection of potential logical gaps, inconsistencies, or misleading expressions
4. Identification of subjective judgments and emotional or opinionated language
5. Overall rationale for the credibility assessment

</think>

Optionally, if applicable, you may also include:

References to commonly known facts or publicly verifiable information. Notes on missing context or insufficient evidence that affect credibility.

Content Veracity: <veracity>[True / False]</veracity>

User Prompt: Please assess the veracity and credibility of the following content. Analyze the text (and image, if provided) step by step, strictly distinguishing factual statements from subjective opinions. Finally, determine whether the content is true or false.

Context: News Text: *[the given news text]*, Image: *[the given news image]*

Figure 10: Label-Only Classification Prompt used by the Qwen2.5-VL-7B SFT baseline. The target sequence contains only the final veracity decision; no reasoning trace and no dimension scores are emitted, so the GSC consistency loss does not apply.