VERIBENCH-WILD: A BENCHMARK AND MEASUREMENT TESTBED FOR MISINFORMATION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce VERIBENCH-WILD, a benchmark and measurement testbed for studying how evaluation choices shape misinformation benchmark scores. A higher score need not mean better detection; it also depends on the information a model receives, its source and adaptation setting, and how its outputs are scored. VERIBENCH-WILD provides 27,158 multilingual but English-dominant social-media records with frozen splits, a primary mechanism-typology task, five further task views, public and full representations, and reproducible reference baselines. Across five evaluated models, removing the fact-check note reduces macro-F1 by 12.7–21.3 points, while note-only inputs remain close to full-input performance. The cross-source gap changes substantially with the adaptation regime, and restricting training and evaluation to common-support examples reduces but does not eliminate it. For generated claims, eight automatic evaluation pipelines, each a judge with its parser, accept the same outputs at rates from 15.3% to 63.4%. These choices change rankings, not only scores: a wrong fact-check note erases the lead of API models over open-weight models, QLoRA fine-tuning beats in-context learning within source but loses to it across sources, and ranking judges by acceptance nearly reverses their ranking by agreement with three blinded human raters. Across all eight judges, the parser matters only when a judge ignores the requested format, which moves one judge from 0.4% to 31.9%. Scores on VERIBENCH-WILD should therefore state their information inputs, source and adaptation settings, and evaluator and parser configuration, so that comparisons are interpreted under explicit conditions.

1 INTRODUCTION

Misinformation detectors are compared by benchmark scores. A higher score is usually read as better detection. Yet two scores on one benchmark can come from different settings. One model may read the fact-check note; another sees only the post. One may be fine-tuned on the training split of the source it is evaluated on; another sees only a few in-context examples. For generated claims, the judge may differ, or the parser that reads its reply. If these settings are not reported, a gap between two scores cannot be attributed to the models. We call the settings behind a score its *measurement contract*: the information the model receives, the sources it is adapted and evaluated on, how it is adapted, and how its outputs become a number.

Prior work covers parts of this problem. Evaluation research shows that one number can hide design choices (Liang et al., 2023; Dehghani et al., 2021); even prompt format can move accuracy (Sclar et al., 2024). Supplied inputs can expose the target (Gururangan et al., 2018; Poliak et al., 2018; Glockner et al., 2022). Success on one source need not transfer (Koh et al., 2021), and transfer depends on how a model is adapted (Mosbach et al., 2023). Automatic judges have their own preferences (Zheng et al., 2023; Wang et al., 2024; Panickssery et al., 2024). These results come from different tasks; we measure such effects on one misinformation resource, holding fixed the prompted model's weights, the initial backbone, or the scored outputs, depending on the audit.

VERIBENCH-WILD serves this purpose. Built from X Community Notes (Wojcik et al., 2022) and from PolitiFact and GossipCop records in UPFD (Dou et al., 2021; Shu et al., 2020), it releases

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.

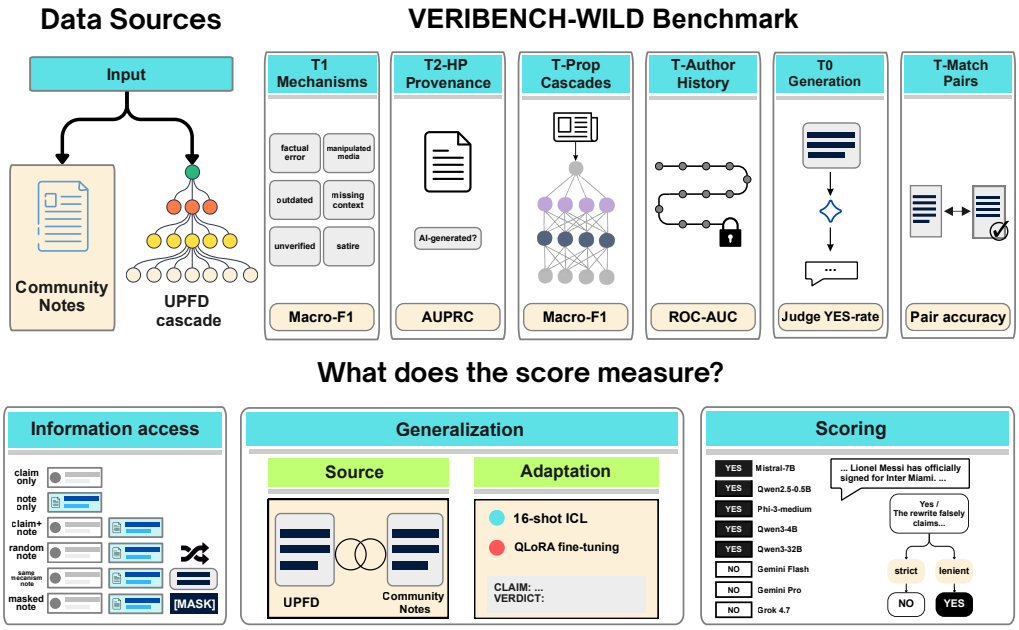


Figure 1: VERIBENCH-WILD overview. Top: two data sources feed six task views, each with its own metric. Bottom: three audits ask what a score measures: information, generalization across source and adaptation, and scoring. They use different tasks and targets, so they are not additive parts of one error measure (results in Figures 2–4; contract schematic in Appendix J).

27,158 records with frozen splits, a primary mechanism-typology task (T1), supporting and diagnostic task views, public and full representations, and reference baselines.[1] Named subsets support three evaluations with different targets (Figure 1): evidence-conditioned mechanism recognition from claims, notes, and the note writers' mechanism tags (T1); binary-verdict transfer between UPFD and Community Notes under stated source pools and adaptation regimes; and scoring of frozen generated claims (T0) by documented judge-and-parser pipelines, checked against independent human annotations. The closest benchmarks center on verdicts, retrieval, social graphs, or multilingual and multimodal coverage (Section 5); these evaluations add mechanism labels, a shared cross-source target, and a human-annotated generation audit. We run them with a fine-tuned encoder, open-weight models (Qwen, Mistral, Phi), and API-only models (Gemini, Grok).

The audits ask three questions, and each finds that the setting changes rankings, not only scores. First, how does permitted information affect T1 scores (Section 4.2)? Removing the note costs every model 12.7–21.3 macro-F1 points, while the note alone comes close to the full input. Notes and tags share an author, so T1 scores reflect evidence-conditioned recognition, not independent verification; this does not show that models ignore the claim. API models outperform open-weight models with a correct, missing, or masked note, but a wrong note erases their lead. Second, how do source pools and adaptation regime affect generalization across sources (Section 4.3)? From one initial backbone, the evaluated QLoRA regime shows a far larger home-source advantage than 16-shot in-context learning, which uses far less supervision: QLoRA wins within source by 17.3 points and loses across sources by 15.2. Common-support matching reduces the QLoRA gap without removing it, but it changes the pools. Third, how do evaluator and parser choices affect generation scores (Section 4.4)? Eight judge-and-parser pipelines differ in acceptance by up to 48 points, and ranking them by acceptance nearly reverses their ranking by agreement with human raters. Checked on all eight judges, the parser matters only when a judge ignores the requested format, which moves one judge from 0.4% to 31.9%.

Our contributions are threefold. (1) *A benchmark resource* (Section 2) with a documented CN-only T1 setup; T2-HP, T-Prop, and T-Author are supporting views, T0 serves the scoring audit, and T-Match is a diagnostic, not a leaderboard task. (2) *Information and source/adaptation audits* (Sections 4.2–4.3): on fixed test records we remove, isolate, swap, or mask the note, and we compare transfer gaps

[1] Code, data, and prompts (anonymized): `https://anonymous.4open.science/r/VeriBench-Wild-ICLR2027-6FE5`.

Task	Input → target	Metric	Reference baseline	Role here
Primary task				
T1	Claim + note → six mechanisms	Macro-F1	Qwen3-4B QLoRA 0.5557	info. audit (§4.2)
Supporting task views (released for reuse)				
T2-HP	Claim → AI-attribution label	AUPRC	Desklib 0.8483	reference
T-Prop	UPFD cascade → verdict	Macro-F1	GraphGPS 0.9683	reference
T-Author	History + time → next attribution	ROC-AUC	History transf. 0.8327	reference
Audit and diagnostic tasks				
T0	Note + mechanisms → new claim	Judge YES-rate	Mistral/Qwen generators	scoring audit (§4.4)
T-Match	Topic + claim pair → selected claim	Pair accuracy	Length rule	diagnostic

Table 1: Task views and their roles. Our conclusions rest on T1, T0, and a binary-verdict transfer setup over named subsets of the same resource (§4.3); the supporting views are released with reference baselines for reuse, and T-Match is a construction diagnostic. Reference scores are seed means on the full view (T2-HP on its strict evaluation pool): Qwen3-4B (Yang et al., 2025) with QLoRA (Dettmers et al., 2023), Desklib (Desklib, 2024), and GraphGPS (Rampášek et al., 2022). Task files, views, and uncertainty are in Appendices A and G.

across source pools and two complete adaptation regimes. (3) *A human-grounded evaluator audit* (Section 4.4): eight judges and three blinded, independent human raters score 476 frozen judgments, and all eight judges' replies are parsed under two rules.

2 The VeriBench-Wild Benchmark Suite

VeriBench-Wild combines X/Community Notes (Wojcik et al., 2022) with PolitiFact and GossipCop records distributed through UPFD (Dou et al., 2021; Shu et al., 2020). Community Notes supplies posts, the notes written about them, and the misleading-reason tags chosen by note writers; the legacy UPFD records supply news claims with propagation graphs. The collection is centered on X/Community Notes and is not a representative sample of all social-media posts or moderation systems.

The frozen release contains 27,158 records: 21,852 training, 2,651 validation, and 2,655 test. Record timestamps end on 2025-02-27. Cleaning standardizes text and label fields, removes duplicate claims, and strips metadata that could leak labels. The main splits keep each author within a single split; author forecasting instead splits each author's history by time.

2.1 Tasks, splits, and reference baselines

T1 is the primary task (Table 1). It trains and selects checkpoints on Community Notes only: 17,728 training, 2,133 validation, and 2,150 test records. UPFD records carry only one of the six mechanism labels, and treating the five missing labels as negatives would teach the model something the data do not say. CN-only training is therefore the official setup; training that also uses UPFD with the missing labels masked can be reported as a separately specified alternative. On the full view, RoBERTa-base (Liu et al., 2019) scores 0.5340 ± 0.0024 macro-F1 over five seeds, and Qwen3-4B QLoRA scores 0.5557 ± 0.0040 on the full view and 0.5542 ± 0.0037 on the public view over three seeds; an earlier pooled-training RoBERTa reference is reported in Appendix D. Architectures and training recipes differ, so these references are comparison points rather than controlled comparisons.

The other views carry construction caveats. T2-HP predicts an AI-attribution label from claim text; the label is derived from the note text by keyword rules (distant supervision), not by human annotation, and does not establish authorship. T-Prop predicts a verdict from legacy UPFD propagation cascades and is close to saturated. T-Author predicts whether an author's next post will be flagged as AI-attributed from up to 16 earlier posts, without the next post's text; because this flag is assigned per collection tier, the task partly amounts to predicting which tier the next post comes from. T0 generates

a claim with a target mechanism from an evidence note, without the original post; its score requires an evaluator and a parser. In T-Match, the correct claim in every pair comes from the supported collection tier, so T-Match diagnoses that construction and is not a leaderboard task.

2.2 Labels and data views

The six T1 labels describe non-exclusive misinformation mechanisms: factual error, manipulated media, outdated information, missing important context, unverified claims presented as fact, and satire. They are the misleading-reason tags that Community Notes writers select when they write a note, so a note and its labels share an author; this matters whenever a model is given the note (Section 4.2). In the original three-rater study on 50 records, which is separate from the generation-judgment study in Section 4.4, Fleiss agreement ranges from $\kappa = 0.080$ for missing context to 0.891 for manipulated media, with a mean of 0.470 (Appendix B). Missing context is also the most frequent label (733 of 2,150 test records), so per-label results for it should be read with caution. Accordingly, T1 evaluates prediction of the released mechanism tags under an evidence-conditioned task contract; it is not an independent oracle of underlying misinformation type.

Each record separates claim text, the evidence note, source metadata, propagation summaries, and labels. Full propagation graphs exist only for the UPFD records; Community Notes records have no retweet trees, and a missing view is never treated as a negative label. The release carries 54 language tags, including codes for undetermined or non-linguistic content; English covers 81.7% of training and 78.9% of test records, so the collection is naturally multilingual but English-dominant.

Access differs by representation (Appendix A, Table 7). The anonymous repository provides the public Tier-1 view, whose three root splits keep claim and note text, task labels, and anonymous author groups but remove post and note identifiers, replace X-internal URLs, and coarsen timestamps; it supports the public T1 references. The repository also holds the frozen prompts, parser rules, aggregate results, and figure script. The full view adds propagation graphs, author histories, pair files, and generation-task files; it is not in the repository, full graphs are gated, and exact author histories are not public. The generation-audit sample, stored judge replies, and individual human annotations of Section 4.4 are not included; their aggregate results are in Appendix F. Scores can differ across views, so new systems train on the official training split, select on validation, and report test results under a stated task setup and view.

3 Overview: The Measurement Contract

We study three parts of the measurement contract on the tasks where they apply (Figure 1): information, generalization across source and adaptation, and scoring. Table 2 summarizes the primary choice under study and the quantities controlled in each comparison. These interventions are task-specific and should not be interpreted as an additive decomposition of benchmark error.

3.1 Information interventions

Supplied evidence is a legitimate T1 input, but a note often explains why a claim is misleading; the question is whether performance under this setup supports claims about independent, pre-evidence verification. We compare six input conditions on the identical 2,150-record CN test pool. The canonical RoBERTa reference uses CN-only training for full, claim-only, and note-only inputs, with a separate checkpoint for each; random-note, same-mechanism wrong-note, and trigger-masked conditions instead modify the full-input checkpoint at inference time. Learned performance with restricted information and sensitivity to evaluation-time corruption therefore remain distinct comparisons. Random notes break the claim–note pairing. Same-mechanism donors share a positive gold label but come from a different record, and records without an eligible donor retain their original note; this is a diagnostic control, not a deployable input strategy. Masking removes a fixed phrase list of explicit misleadingness cues, so it tests those phrases, not all lexical information in the evidence.

3.2 Generalization: source and adaptation

A cross-source gap can reflect the collections, the adaptation regime, and their interaction. We therefore hold the initial Qwen3-4B-Instruct-2507 backbone fixed while comparing 16-shot in-context

Choice	Task and pool	What varies	Controlled	Models
Information	T1; 2,150 CN test records	Claim only, note only, claim + note; random, same-mechanism, or masked note	Test records and labels; prompted-model weights	RoBERTa (fine-tuned); Qwen3-4B, Qwen3-32B (open); Gemini Flash, Gemini Pro (API)
Source	Binary verdict; UPFD vs. CN, unmatched and strict-matched pools	Source used for adaptation and for testing	Initial backbone, claim-only input, target, scoring	Qwen3-4B
Adaptation	Same pools as the source row	16-shot in-context learning vs. QLoRA fine-tuning	Initial backbone, evaluation records	Qwen3-4B
Scoring	T0; 384 items, 476 judgments	Judge (eight); parser (strict vs. lenient)	Generated outputs, prompt, mechanism definitions	Mistral-7B, Qwen2.5-0.5B, Phi-3-medium, Qwen3-4B, Qwen3-32B (open); Gemini Flash, Gemini Pro, Grok 4.7 (API); three human raters

Table 2: Audit design. Each row identifies the primary contract choice under study and the quantities controlled in that comparison. The source and adaptation rows form the generalization audit (§4.3). In the information audit, full, claim-only, and note-only RoBERTa use separately trained checkpoints; in the adaptation audit, QLoRA and in-context learning also differ in the amount of supervision.

learning (ICL) with source-specific QLoRA (Dettmers et al., 2023). UPFD (U) and Community Notes (C) share a binary SUPPORTED/REFUTED target for this audit; inputs are claim text only, and the target is distinct from T1's six mechanisms. Within each unmatched or strict-matched pool, both regimes use the same evaluation records and score the same two label continuations by summed log probability, so no generated-answer parser enters this comparison. ICL draws 16 label-balanced demonstrations from one source's training records; QLoRA trains on that source's available training split. We evaluate within and across sources with three demonstration or training seeds. For source-conditioned macro-F1 $F_{s\to t}$, the bidirectional home-source advantage in percentage points is

$$G = 100\,\frac{(F_{U\to U} - F_{U\to C}) + (F_{C\to C} - F_{C\to U})}{2}. \tag{1}$$

A small G need not mean high accuracy, since it also arises when within-source and cross-source scores are both modest; complete four-cell matrices therefore accompany the gaps (Figure 3; Appendix E). Matching restricts observable common support using language, binary verdict, claim length, and frozen embedding topics, while preserving original split membership. The strict pool retains 1,802 UPFD and 1,400 CN records across splits. This changes both training and evaluation sets; it does not hold sample size or test difficulty fixed.

3.3 Evaluation: judges, parsers, and human annotation

A generation score depends on a judge's decisions and on how they are extracted; we examine both on the same frozen outputs and compare them with independent human annotations. The audit contains 384 generated items from Mistral and Qwen systems, yielding 476 mechanism judgments (240 for Mistral generations and 236 for Qwen), because multi-mechanism items contribute several judgments. Sampling balances generator/mechanism strata, with a separate multi-mechanism stratum; the sample is not a random draw from all possible generations. Accordingly, the reported acceptance rates characterize this audit sample rather than unrestricted generator success rates. The eight judges in Table 2 evaluate this pool under the specified prompt and mechanism definitions. They differ in scale, training, and inference settings, so the audit measures evaluator choice as a whole (identifiers and protocol in Appendix H). The frozen strict parser accepts only responses with the requested `VERDICT:` field; a lenient parser also recognizes a bare leading verdict. The original three judges stored only parsed verdicts, so we re-ran them under the frozen protocol to store raw replies for both parsers. All 476 judgments stay in the denominator, and an unparsed response counts as not-YES.

Three blinded, independent raters each annotated all 476 judgments using the same interface, batch, and decision rules, with independently randomized item orders. They used the automatic evaluators' mechanism definitions and did not see generator or automatic-judge identities. Human-majority YES requires at least two YES votes; other response labels count as not-YES in the binary comparison.

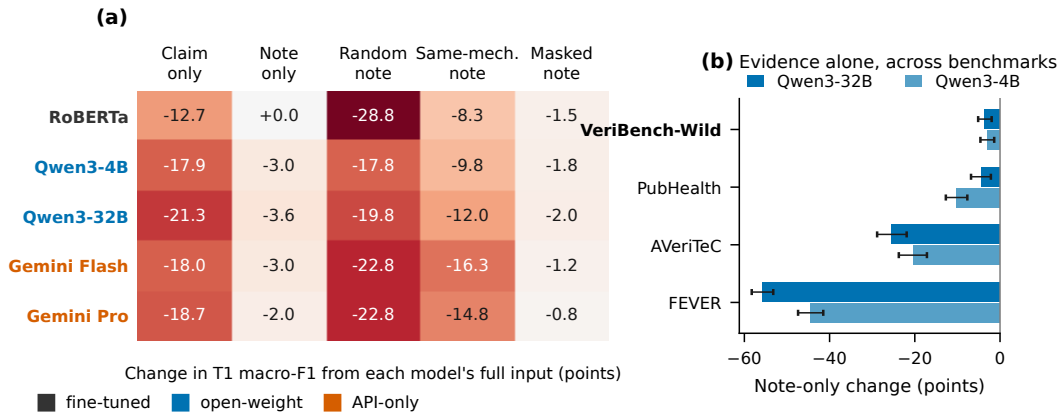


Figure 2: Information access. (a) Change in T1 macro-F1 from each model's full input on the same 2,150 Community Notes test records, under five input conditions. RoBERTa is fine-tuned (paired-t over five seeds); the other models are prompted zero-shot (paired item bootstrap). (b) Change when only the evidence is given, for Qwen3-32B and Qwen3-4B on four benchmarks, each within its own label space (95% paired intervals; Appendix D.1).

4 Characterizing What the Scores Measure

4.1 Experimental setup

We evaluate three kinds of widely used models (Table 2). The fine-tuned encoder is RoBERTa-base (Liu et al., 2019), trained on T1 with five seeds. Freely available open-weight models are Qwen3-4B and Qwen3-32B (Yang et al., 2025), used zero-shot on T1 and as judges, with Qwen3-4B also the fixed backbone of the adaptation audit; Mistral-7B (Jiang et al., 2023), Qwen2.5-0.5B (Yang et al., 2024), and Phi-3-medium (Abdin et al., 2024) serve as judges. API-only models, whose weights are not accessible, are Gemini-3.5-Flash and Gemini-3.1-Pro (T1 and judging) and Grok-4.7 (judging). Figure 2 compares changes within each model, avoiding absolute-level comparisons between fine-tuned classification and prompted inference. Uncertainty follows each design: paired differences across five seeds for RoBERTa, paired item bootstraps for the fixed prompted models, t-intervals over three demonstration or training seeds for adaptation, and an item-clustered bootstrap for T0.

4.2 Scores depend on the fact-check note across model types

Removing the note costs every model a large share of its score, while the note alone is nearly enough (Figure 2a). For CN-only RoBERTa, claim-only input loses 12.71 macro-F1 points, with 95% paired-t interval $[-13.48, -11.94]$. Note-only differs from full input by just $+0.04$ points, with interval $[-0.35, +0.44]$; we claim neither superiority nor equivalence. Removing evidence costs 17.9–21.3 points in the four prompted models, while their note-only deficits are only 2.0–3.6 points. Random notes cost 17.8–28.8 points across the five models and same-mechanism wrong notes 8.3–16.3 points, whereas masking the fixed trigger list costs at most 2.0 points. Because the T1 labels are the note writers' own tags (Section 2), a close note-only score is expected. The common pattern supports evidence-conditioned mechanism recognition, without showing that models retrieve evidence or verify claims independently. Per-condition tables are in Appendix D.

The evidence condition also decides which models look better (Appendix K). With the correct note, both API models outperform both open-weight models by 2.2–5.6 points, and all paired intervals exclude zero; the mean API lead is 3.9 $[2.5, 5.2]$ points and persists with the note removed (5.1 $[3.5, 6.7]$) or masked. A wrong note erases it: the lead is -0.1 $[-1.7, 1.4]$ with a random note and -0.8 $[-2.2, 0.7]$ with a same-mechanism note, whose point estimates reverse the model order. API models thus lose 4.0–4.6 points more than open-weight models when the evidence is wrong. An external diagnostic shows that explanatory evidence can carry the target: with Qwen3-32B, evidence-only deficits are 3.6 points on VeriBench-Wild and 4.5 on PubHealth (Kotonya & Toni, 2020), against 25.4 on AVeriTeC and 55.8 on FEVER (Figure 2b; Appendix D.1), although tasks, evidence selection, and truncation differ. A T1 result should therefore name its permitted input channels

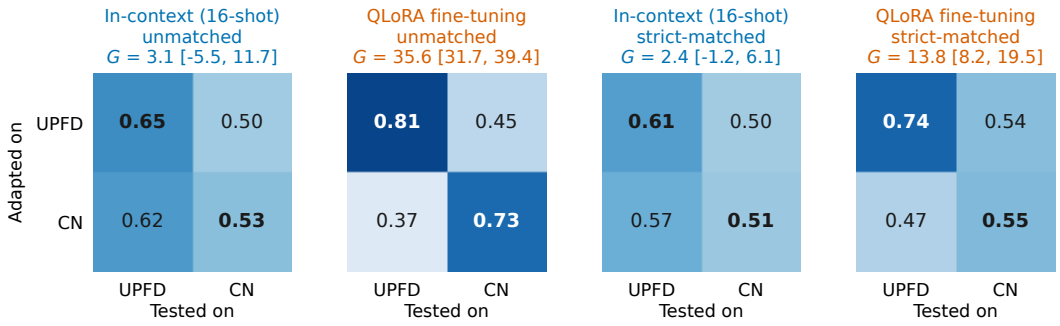


Figure 3: Source and adaptation on the same initial Qwen3-4B backbone. Each 2×2 panel shows binary-verdict macro-F1 when adapted on one source (rows) and tested on each source (columns), averaged over three seeds; diagonal cells are home-source scores. G is the bidirectional home-source advantage of Equation 1 with its 95% t-interval. Four-cell tables with standard deviations are in Appendix E.

and evidence condition, which can reverse a model ranking, and be read as evidence-conditioned mechanism recognition, not independent claim verification.

4.3 The cross-source gap depends on the adaptation regime

With the initial backbone and evaluation pools fixed, the evaluated QLoRA regime exhibits a strong home-source advantage, whereas the 16-shot ICL interval includes zero (Figure 3). On unmatched pools, 16-shot ICL has $G = 3.07$ points, with 95% interval $[-5.51, 11.66]$, while QLoRA has $G = 35.56$ $[31.67, 39.45]$. Which regime looks better depends on where it is tested: on unmatched pools, QLoRA beats ICL within source by 17.3 $[14.3, 20.3]$ points but trails it across sources by 15.2 $[8.6, 21.7]$ points (seed-paired t-intervals; Appendix K).

Under strict matching, the QLoRA gap falls from 35.56 to 13.83 points $[8.21, 19.45]$, while ICL has $G = 2.43$ $[-1.20, 6.06]$. Arithmetically, the 21.73-point reduction combines a 12.37-point drop in mean within-source macro-F1 (0.7666 to 0.6429) with a 9.36-point rise in mean cross-source macro-F1 (0.4110 to 0.5046; Appendix E). Because matching changes training size, composition, and evaluation examples together, this is an accounting of the reported cells, not a causal decomposition. An earlier harmonized RoBERTa comparison gives supporting context (Appendix G.5).

This comparison does not isolate adaptation mechanism: QLoRA uses many more source-labeled examples and updates weights, ICL conditions on 16 examples, and numerical precision differs (a zero-shot NF4 control accompanies the bfloat16 baseline). Zero-shot Qwen already differs between UPFD and CN, and CN-trained QLoRA predicts SUPPORTED for only about 5–9% of unmatched UPFD examples against 55.7% in the gold labels, which is consistent with source-conditioned decisions but does not prove a learned source prior. A transfer result should therefore state its source pools, matching rule, adaptation regime, and supervision budget, and report all four source-to-target scores beside G, since a smaller gap can accompany lower within-source performance.

4.4 Judges and parsers change what counts as a success

Across the eight judge-and-parser pipelines, acceptance of identical outputs ranges from 15.3% to 63.4% (Figure 4a; Table 3). The spread holds within each scoring convention: 33.0 points among the three original judges (stored strict verdicts) and 16.6 among the five added judges (lenient parser). Every automatic judge accepts Mistral generations more often than Qwen generations, by 4.0 to 17.1 points (Figure 4b), which does not by itself identify generator quality or evaluator self-preference.

Human-majority acceptance is 16.4%: 19.2% for Mistral generations and 13.6% for Qwen. The 5.6-point difference has an item-clustered 95% interval of $[-0.6, 12.2]$, so humans show the same direction but the interval includes zero. Human agreement is moderate and construct-dependent (Table 4): Fleiss κ is 0.534 (four labels) and 0.507 (YES vs. not-YES), and binary agreement ranges from 0.03 (outdated) to 0.73 (unverified). Acceptance is not agreement: ranking the eight

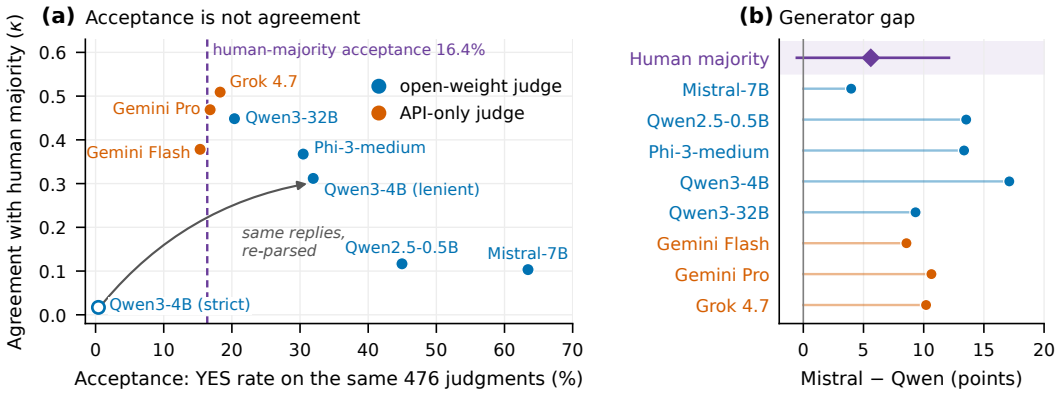


Figure 4: Scoring the same 476 T0 judgments. (a) Acceptance (YES rate) against agreement with the human majority (Cohen's κ). The dashed line marks human-majority acceptance; the arrow joins Qwen3-4B's stored replies read by the strict and by the lenient parser. (b) Difference between acceptance of Mistral and of Qwen generations, with the item-clustered 95% interval for the human majority. Original judges (Mistral-7B, Qwen2.5-0.5B, Phi-3-medium) use stored strict verdicts and added judges the lenient parser; rows follow the audit's fixed order, not a ranking.

Judge	Type	YES%	κ	BA	F1
Mistral-7B†	open	63.4	0.103	0.611	0.337
Qwen2.5-0.5B†	open	45.0	0.117	0.599	0.329
Phi-3-medium†	open	30.5	0.367	0.747	0.502
Qwen3-4B	open	31.9	0.312	0.715	0.461
Qwen3-32B	open	20.4	0.448	0.746	0.549
Gemini Flash	API	15.3	0.378	0.684	0.477
Gemini Pro	API	16.8	0.469	0.737	0.557
Grok 4.7	API	18.3	0.509	0.766	0.594
Qwen3-4B, strict	open	0.4	0.017	0.505	0.025
Human majority	—	16.4	—	—	—

Table 3: Judges vs. the human majority on 476 judgments (fixed order, not a ranking). †Stored strict verdicts; others use the lenient parser. BA: balanced accuracy; F1: YES-class F1.

Mechanism	n	κ_2	κ_4
Factual error	93	0.383	0.495
Manipulated media	69	0.328	0.349
Outdated	78	0.034	0.085
Missing context	89	0.238	0.228
Unverified claim	85	0.726	0.712
Satire	62	0.565	0.455
All judgments	476	0.507	0.534

Table 4: Agreement among the three human raters (Fleiss κ) by target mechanism: κ_2 for YES versus not-YES, κ_4 for the four response labels. n counts mechanism judgments.

judges by acceptance nearly reverses their ranking by agreement with the human majority (Spearman $\rho = -0.81$; Appendix K), so the most accepting judges agree least (Figure 4a).

We applied both parsers to all eight judges, re-running the three original judges under the frozen protocol to store their raw replies (their strict verdicts reproduce 93.9–100% of the stored ones; Appendix K). The parser matters selectively: seven judges change by at most 0.2 points, because they follow the requested template or leave it mainly with bare NO replies. Qwen3-4B follows the template in 1.1% of replies and moves from 0.4% to 31.9% YES, and from last to sixth in agreement with the human majority. The frozen prompt asks for exactly one verdict token but also specifies a `VERDICT:`-prefixed format (Appendix H), so this is protocol-specific response-format sensitivity, not worse judgment. It still changes a conclusion across judges: under strict parsing, the acceptance–agreement rank correlation falls from -0.81 to -0.14. Generators see only the note and target mechanisms, whereas the evaluator also sees the original claim and judges each output as a rewrite of it; the scoring results characterize this protocol. A generation score should therefore report the judge and its version, prompt, parser, and denominator, with agreement against human labels where available, because acceptance alone is neither correctness nor agreement.

5 Related Work

Table 5 compares VeriBench-Wild with the closest released resources. CommunityFact is closest in source: it extracts true/false claims from noted posts, keeps note URLs as evidence leads,

Resource	Main task or target	Released data and context	Relation to this work
AVeriTeC	Verdicts on real-world fact-checked claims	4,568 claims from 50 organizations; question–answer web evidence	Evidence-only comparison (Fig. 2b)
FakeNewsNet, UPFD	Fake vs. real PolitiFact and GossipCop news	Social context; propagation graphs with user-history features	Source of our UPFD records
MuMiN	Claim and tweet veracity (node classification)	21M tweets linked to 13K claims; 41 languages; images	Broader social and image context
CommunityFact	True/false claims from posts with Community Notes	15,992 text claims; five languages, two domains; refreshable	Same platform; verification, not mechanism tags
VeriTaS	Multimodal fact-checking, standardized verdicts	25,000 claims; 54 languages; audiovisual; quarterly updates	Dynamic; ours is a static snapshot
VERIBENCH-WILD	Mechanism typology (T1); further task views	27,158 frozen CN and UPFD records; writer tags; public and full views	Controlled audits (§4)

Table 5: The closest released resources (Schlichtkrull et al., 2023; Shu et al., 2020; Dou et al., 2021; Nielsen & McConville, 2022; Singh et al., 2026; Rothermel et al., 2026). Middle columns describe tasks and released data as reported in each paper, not experiments.

and its accompanying study tests web-search LLMs (Singh et al., 2026). LIAR and FEVER are standard verdict references (Wang, 2017; Thorne et al., 2018), and Community Notes research studies participation and helpfulness (Pröllochs, 2022; Xing et al., 2026). These resources offer retrieval, updates, more languages, or media that our static, text-based, English-dominant snapshot lacks.

VERIBENCH-WILD combines notes with their writers' mechanism tags, two sources with a shared verdict target, and generation-task files with controlled audits on named subsets, including a human-annotated generation sample. The idea that evaluation depends on protocol is not new. Our information audit follows hypothesis-only and annotation-artifact analyses (Poliak et al., 2018; Gururangan et al., 2018; McCoy et al., 2019; Schuster et al., 2019); here, the evidence is a legitimate input. Glockner et al. (2022) show that fact-checking datasets often supply counter-evidence leaked from fact-checking articles, unavailable when a claim first spreads. Community Notes play this role in T1; our note-only condition measures how much of the task they carry. Our source audit follows work on distribution shift (Koh et al., 2021; Dehghani et al., 2021), including the MISINFO-GENERAL benchmark of temporal, topical, publisher, and other misinformation shifts (Verhoeven et al., 2026); ours tracks one named gap under matching and adaptation. Our evaluator audit follows studies of LLM judges and self-preference (Zheng et al., 2023; Liu et al., 2023; Panickssery et al., 2024; Chen et al., 2025). Its human annotations are a reference for agreement, not an oracle. Our guidance parallels prompt-format spread (Sclar et al., 2024) and multi-metric evaluation (Liang et al., 2023).

6 DISCUSSION

The reporting items closing Sections 4.2–4.4, with the data view, CN-only T1 setup, and T0 denominator, form a reporting contract for results on the benchmark.

The collection is Community Notes-centered, English-dominant, and shaped by accessible posts and legacy fact-check selection, so it does not establish generalization to all platforms. Media-related labels can lack the underlying media, propagation and history views have unequal source coverage, typology labels remain ambiguous, and human generation agreement varies strongly by mechanism. The static snapshot cannot rule out training exposure of recent models. Matching leaves unobserved source differences, QLoRA versus ICL changes the amount and form of supervision as well as the adaptation mechanism, external evidence regimes are illustrative, and parser failures are selective.

7 CONCLUSION

VERIBENCH-WILD provides frozen splits and task views for mechanism recognition, cross-source generalization, and generation scoring. On it, every audited setting changed model rankings, not only scores. A wrong fact-check note erased the lead of API models over open-weight models, QLoRA beat ICL within source but lost to it across sources, and judges that accepted more agreed less with human raters. Results should state these settings and fix them when comparing systems.

Ethics and Data Statement

The data comes from public posts, Community Notes, and legacy releases. Public availability does not remove privacy or linkage risks. The public distilled view removes direct identifiers and coarsens timestamps, but retained text can remain searchable. Full graph access is gated. Individual poster consent was not obtained, and no ethics-board filing was made. Generation is intended for controlled evaluation and can be misused to create misleading content. Appendix I discusses access, annotation, and use limitations.

AI Use Statement

Generative AI tools were used to assist with manuscript organization, language polishing, and presentation. All experiments, quantitative analyses, scientific decisions, and final claims were conducted and verified by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Reproducibility Statement

The appendices specify task files, training and scoring protocols, full/public references, uncertainty calculations, and exact prompts. All data figures and the numerical appendix tables are generated from a hashed extract of stored results by the released figure scripts. No model calls are needed to rebuild the paper. The anonymous repository linked in the introduction contains the public Tier-1 splits, the frozen prompts and parser rules, the aggregate result extract, and the script that regenerates Figures 2–4 and Tables 3–4. Public Tier-1 supports the reported T1 references; additional task-specific inputs remain necessary for other views.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, 2025. doi: 10.18653/v1/2025.emnlp-main.86. URL https://aclanthology.org/2025.emnlp-main.86/.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021. URL https://arxiv.org/abs/2107.07002.

Desklib. Desklib AI-generated text detector v1.01. HuggingFace model desklib/ai-text-detector-v1.01, 2024. Trained on the RAID benchmark (Dugan et al., 2024). Accessed 2026-05.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yingtong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 2051–2055, 2021.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723.

Max Glockner, Yufang Hou, and Iryna Gurevych. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5916–5936, 2022.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT, Short Papers*, pp. 107–112, 2018.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023.

Pang Wei Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3428–3448, 2019.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12284–12314, 2023.

Dan Saattrup Nielsen and Ryan McConville. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 3141–3153, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 180–191, 2018.

Nicolas Pröllochs. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 794–805, 2022. doi: 10.1609/icwsm.v16i1.19335.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mark Rothermel, Marcus Kornmann, Marcus Rohrbach, and Anna Rohrbach. VeriTaS: The first dynamic benchmark for multimodal automated fact-checking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42066–42100, 2026. doi: 10.18653/v1/2026.acl-long.1948. URL https://aclanthology.org/2026.acl-long.1948/.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3419–

3425, 2019.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020. doi: 10.1089/big.2020.0062.

Sahajpreet Singh, Insyirah Mujtahid, Min-Yen Kan, and Kokil Jaidka. CommunityFact: A dynamic, multilingual, multi-domain benchmark for misinformation detection in the wild. arXiv preprint arXiv:2605.30241, 2026. URL https://arxiv.org/abs/2605.30241.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pp. 809–819, 2018.

Ivo Verhoeven, Pushkar Mishra, and Ekaterina Shutova. Yesterday's news: Benchmarking multi-dimensional out-of-distribution generalization of misinformation detection models. *Computational Linguistics*, 52(2):619–675, 2026. doi: 10.1162/coli.a.585. URL https://aclanthology.org/2026.cl-2.6/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

William Yang Wang. "Liar, Liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pp. 422–426, 2017.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M.B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, 2022.

Rui Xing, Preslav Nakov, Timothy Baldwin, and Jey Han Lau. COMMUNITYNOTES: A dataset for exploring the helpfulness of fact-checking explanations. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1390–1411, 2026. doi: 10.18653/v1/2026.findings-eacl.71. URL https://aclanthology.org/2026.findings-eacl.71/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, et al. Qwen2.5 Technical Report, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

A Dataset Details and Task Files

The release follows a standardized record schema and supplies task-specific views. This appendix separates the full-text scientific inputs from the public distilled representation. Their records correspond, but URL redaction can change tokenization and therefore model predictions.

Composition. The root files contain 21,852 training, 2,651 validation, and 2,655 test records. The total is 27,158. Source tiers contain 5,147 UPFD legacy records, 5,262 `wild`, 8,701 `wild_control`, and 8,048 `wild_support` records. These are collection strata, not interchangeable ground-truth categories. The release contains 54 language tags, and its latest recorded post timestamp is 2025-02-27. The language distribution is English-dominant (81.7% train; 78.9% test). A release build date is distinct from a record's creation date.

Construction and missing views. The build joins Community Notes annotations to accessible post text and standardizes legacy records. It collapses multiple notes, normalizes schema fields, removes duplicates, and hashes author handles. The frozen release also removes selected source/tier and cited-URL fields from evaluation inputs to restrict leakage. These exclusions are task-specific; provenance fields retained for audit must not silently enter a classifier's feature set. Community Notes provides evidence text and engagement summaries, not equivalent legacy cascade graphs. The text release does not include the original image, video, or audio bytes.

View	Research content	Boundary
Claim	Normalized post or legacy claim text	May reference unavailable media
Evidence	Note text and available evidence metadata	Availability and permitted use vary by task
Source	Provenance and pseudonymous grouping fields	Audit metadata is not automatically a model input
Propagation	Legacy graphs or available summaries	Full graphs require the gated view
Labels	Mechanisms, verdict, AI-attribution and other released targets	Different tasks use different label subsets
Temporal	Record times and history ordering	Public view coarsens exact timestamps

Table 6: Schema-level views. Field availability is explicit; missing contextual information must not be interpreted as evidence of truthfulness.

Official file contracts. Paths below are relative to the full release's `data/v4_1_clean_full/` directory. Root splits are `veribench_wild_train.jsonl`, `veribench_wild_val.jsonl`, and `veribench_wild_test.jsonl`. Integrity information is stored in `SHA256_MANIFEST.txt`.

- **T1:** root splits with six `labels.sub_reasons` mechanism targets. The official training, validation, and test partitions restrict to Community Notes (17,728/2,133/2,150 records). Pooled training is historical robustness only; unavailable UPFD labels are not canonical negative targets.
- **T2-HP:** `media_ai_attribution/`; strict evaluation retains `L_HIGH` positives and `NEG` negatives only. The classifier uses claim text, never `evidence.note_text`. The strict test has 436 positives and 748 negatives.
- **T-Prop:** `upfd_legacy_only/`; binary SUPPORTED/REFUTED. Train/validation/test files contain 4,132/518/505 rows. Their total includes eight documented upstream cross-listings, so it differs from the unique legacy-record total.
- **T-Author:** `author_trajectory/`; use the task's per-author temporal ordering. A history contains at most 16 prior observations. Target time supplies a time-delta feature, but target claim text is excluded.
- **T0:** `counterfactual/`; the historical directory name identifies the released files. The operational task maps note text and target mechanisms to a generated claim exhibiting them. It is not an edit of an input original claim.

- **T-Match:** `topic_matched/`; same-topic pairs form a construction diagnostic. The collection tier and pair target are confounded, so pair accuracy should not be treated as a general veracity leaderboard.

Splits and audit pools. Primary author-disjoint splits prevent shared author groups across partitions. The author-history task intentionally retains authors across time partitions. Information auditing reuses the 2,150 non-legacy test records with evidence notes. Source matching retains record IDs within their existing split assignments; it does not randomly repartition matched records. The evaluator audit samples stored generations from the generation task, not new real-world claims.

Public and gated access. The distilled view lives in `data/v4_1_tier1_distilled/`, with `train.tier1.jsonl`, `val.tier1.jsonl`, `test.tier1.jsonl`, and `SHA256SUMS`. It retains claim/note text, task labels, anonymous author grouping, and aggregate cascade statistics. It removes direct post and note identifiers, raw/hashed handles, and exact timestamps. X-internal URLs are replaced by a placeholder. Full graph sidecars are not shipped in this public view. Retained text can remain searchable, so this is linkage reduction rather than a guarantee of anonymity.

Historical experiments use full-text inputs; new T1 references additionally evaluate Tier-1. Exact reproduction requires those inputs and the saved task-specific history features or timestamps where applicable. Rebuilding a history from month-bucketed root records alone need not reproduce the original time-delta features. Public-view runs should therefore identify their representation and report their own reference scores. This distinction preserves a usable public benchmark without claiming bit-identical results for changed inputs.

Three representations and task compatibility. The full scientific view, public full-layout mirror (`data/v4_clean_full_public/`), and distilled Tier-1 view are distinct. The mirror retains selected sidecars, whereas Tier-1 ships only the three root splits. The new Qwen3-4B QLoRA CN-only reference is evaluated on both full and Tier-1 views; historical pooled public references are separately labeled. CN-only public RoBERTa was not run. Researchers must state the representation, permitted fields, and split used rather than assume identical scores across views.

Task / audit	Tier-1 compatibility	Additional requirement or boundary
T1 / information	Protocol constructible from claim, note, and labels	Redaction changes tokens; fixed trigger list belongs to experiment code
Source transfer	Protocol constructible from text and source/verdict fields	Redaction changes length and topic features, hence matching
T2-HP strict	Strict labels absent	`media_ai_attribution/` sidecar; available in the full-layout mirror, not Tier-1
T-Prop GraphGPS	Summary statistics only	Full/gated graph structure, not just a legacy-row sidecar
T-Author	Exact time/history features absent	Full `author_trajectory/`; month buckets cannot recover time deltas
T-Match	Pair files absent	Full `topic_matched/`; external probes use their own corpora
T0 generation	Generation sidecar absent	`counterfactual/` in full or public full-layout view, plus model/prompt contract
Evaluator audit	Experiment pool absent	Frozen generated-item pool and judge protocol, separate from the benchmark release

Table 7: Representation-specific reuse contract. Protocol compatibility is not bit-identical reproduction. Full-layout sidecar availability does not imply that gated graphs or exact author histories are public.

Documentation and maintenance. The release supplies manifests and data documentation, following the purpose of datasheets and data statements (Gebru et al., 2021; Bender & Friedman, 2018). The anonymous review repository is linked in the introduction. Access arrangements, license files, and removal procedures should accompany any redistribution. No future archive identifier, maintenance deadline, or institutional approval is asserted here.

B Annotation Guidelines and Agreement

The typology labels originate in Community Notes. This appendix describes the original dataset-label studies, including the 50-record typology study. The separate generation-validation study uses three independent raters on all 476 T0 judgments (Appendix F). The studies assess different annotation targets and do not relabel the whole corpus.

Study design. The source study reports four PhD-level annotators on 190 binary AI-attribution records and three trained annotators on 50 typology records. The binary study reports Fleiss $\kappa = 0.94$. The typology study reports a mean Fleiss κ of 0.47, with substantial variation across mechanisms. Recruitment came from the authors' research network rather than paid crowd work. These small studies are not representative multilingual validation.

Instructions. Annotators read the claim and Community Notes rationale. They could mark each mechanism YES, NO, or SKIP, and multiple mechanisms could apply. YES means the mechanism applies; NO means the rater confidently judged it absent. SKIP records ambiguity, insufficient context, or a language barrier. Raters worked independently and did not see each other's decisions. The rationale was explicitly described as a primary signal for why the post was flagged.

Mechanism	Annotation definition	Fleiss κ
Factual error	A verifiable false factual assertion	0.401
Manipulated media	Fake, edited, AI-generated, or misattributed media	0.891
Outdated information	Previously true information presented without its changed temporal context	0.464
Missing important context	Omitted context materially changes the claim's meaning	0.080
Unverified claim as fact	An allegation or unsubstantiated claim presented as established fact	0.380
Satire	Satirical content presented or interpreted as serious	0.603

Table 8: Typology annotation instructions, paraphrased, and three-rater agreement on 50 records. The mean across mechanisms is 0.470. These are agreement statistics, not accuracy estimates or performance ceilings.

Interpretive limits. Reading an explicit AI-attribution phrase can make agreement easier without providing independent forensic evidence about authorship. Missing-context decisions are harder to operationalize: raters may disagree about whether any relevant omission qualifies or only an omission that changes meaning. Low agreement can reflect ambiguity, prevalence, and rater differences. It should not automatically be assigned to model undertraining or to a defective statistic.

The original 50-record typology study uses the dataset annotation rubric above, which differs in places from the frozen T0 evaluator definitions (Appendix H). Its annotation instructions and browser interface are retained with that study's packet. In the separate 476-judgment generation-validation study, all three human raters use the same target-mechanism definitions as the automatic T0 evaluators. Agreement from the original typology study therefore remains distinct from the direct human–judge comparison on generated items.

C Reference Recipes and Reproduction Protocol

Reference baselines provide starting points under fixed task contracts. They are not claims of state-of-the-art performance. The frozen scripts and run configurations determine reproduction details; the compact recipes below identify the principal settings.

Task / model	Optimization	Input and model settings
T1 RoBERTa	LR $2\,10^{-5}$; batch 16; 3 epochs	256 tokens; six-label binary cross-entropy; validation macro-F1 selection
T2-HP RoBERTa / Desklib	LR 10^{-5}; batch 16, accumulation 2; 4 epochs	Claim-only, 512 tokens; class-weighted binary objective; validation AUPRC selection
T0 Mistral	LR $2\,10^{-4}$; batch 4, accumulation 8; 3 epochs	QLoRA rank 16, scale 32, dropout 0.05; 2,048 training tokens
T0 Qwen	LR $2\,10^{-5}$; batch 8, accumulation 2; 3 epochs	Full fine-tuning; 1,024 training tokens; four available seeds
T-Prop GraphGPS	LR $3\,10^{-4}$; batch 32, accumulation 2; 30 epochs	Four GPS layers; hidden size 128; four attention heads; binary verdict
T-Author transformer	LR $3\,10^{-4}$; batch 64; 8 epochs	Four layers, hidden size 256, eight heads; frozen MiniLM history embeddings; history length 16
Source audit RoBERTa	LR $2\,10^{-5}$; batch 32; 3 epochs	Claim-only, 256 tokens; same recipe for harmonized unmatched and matched settings

Table 9: Reference training recipes. LR denotes learning rate. Task scripts and run configurations remain the detailed authority; RoBERTa audits use five seeds; modern supervised references and same-backbone source experiments use three.

Reference implementations. RoBERTa follows Liu et al. (2019). T0 uses Mistral and Qwen backbones (Jiang et al., 2023; Yang et al., 2024), with QLoRA for Mistral (Dettmers et al., 2023). GraphGPS provides the graph reference (Rampášek et al., 2022). The author-history transformer consumes embeddings and history statistics, not the next claim's text. A simple recent attribution rate and a logistic model over history statistics are required comparison points.

Input processing. The RoBERTa information audit serializes full input as claim, a separator, and note, truncated from the end at 256 tokens. Claim-only and note-only conditions use their corresponding trained checkpoints. Random-note, same-mechanism donor, and trigger-mask conditions retain the full-input checkpoint. Donor assignment is fixed across model seeds. Trigger masking is case-insensitive over a saved phrase list, rather than an adaptive search for model-sensitive tokens.

Additional optimization settings. The T1 RoBERTa configuration uses float32, weight decay 0.01, linear scheduling, warmup ratio 0.1, and no label smoothing. Its binary cross-entropy uses training-derived positive-class weights capped at 20. The T2-HP trainable references use bfloat16, positive weight 5.9, label smoothing 0.01, dropout 0.1, weight decay 0.01, and linear scheduling with warmup ratio 0.06. Their validation-AUPRC stopping patience is two evaluations, with minimum improvement 0.0001.

Mistral uses NF4 double quantization with bfloat16 computation and adapters on the query, key, value, output, gate, up, and down projections. Both generators use cosine scheduling, warmup ratio 0.03, and weight decay 0.01. The saved Mistral configuration names the last-epoch checkpoint; Qwen names the best validation-loss checkpoint. Their configuration files also record validation-loss stopping parameters, so a common checkpoint-selection rule should not be assumed.

GraphGPS uses balanced cross-entropy, label smoothing 0.05, dropout 0.1, gradient clipping at norm 1.0, weight decay 0.01, and cosine scheduling with warmup ratio 0.1. It uses eight Laplacian positional components and 16 random-walk steps, with stopping patience eight. The history transformer uses a feedforward size of 1,024 and input dimension 389. Its configuration uses label smoothing 0.02, weight decay 0.01, cosine scheduling with warmup ratio 0.06, and validation-ROC-AUC stopping patience three.

Source-audit implementation boundary. All harmonized RoBERTa source conditions use the matched runner's training and evaluation implementation. The harmonized unmatched control changes only the record-ID source and output directory. Training uses three epochs, batch size 32, maximum length 256, and constant learning rate $2\,10^{-5}$. AdamW weight decay is 0.01. Models train

only on training records and evaluate the final checkpoint on the validation-plus-test union; that union is not used for checkpoint selection. The historical unmatched runner instead used linear decay with 10% warmup and different shuffling details. Appendix G.5 separates that historical comparison from the common-runner result.

Seeds and uncertainty. The standard seed set is $\{7, 13, 42, 43, 44\}$. Mistral generation has five stored seeds; the Qwen generation reference has $\{7, 13, 43, 44\}$, with seed 42 absent. Deterministic detector inference has one stored seed. A filename containing `5seed` does not by itself establish five independent stochastic runs. Repeated deterministic results are recorded as such.

The canonical RoBERTa information audit uses paired-t intervals over five seed differences. Appendix G.2 retains 10,000-resample paired-seed percentile intervals for continuity, alongside raw values and paired-t sensitivity intervals. The historical RoBERTa source audit follows the same five-seed reporting discipline. The new Qwen source audit uses three-seed t intervals (Appendix E); prompted information replications use paired item intervals, and human comparisons use the item-clustered procedure in Appendix F. Item-level intervals for a fixed model instead resample evaluation records. The evaluator difference-in-differences uses 10,000 item resamples, and its regression interval uses 500 item-clustered resamples. These intervals answer different uncertainty questions. They do not quantify uncertainty over future platforms or all possible benchmark constructions.

Compute accounting. The historical runs used H100 80GB GPUs, with CPU execution for applicable statistical baselines. A complete independently traceable total GPU-hour or carbon ledger is not available in the frozen reporting audit. We therefore omit a numerical total. Reproducing the paper requires model training or saved predictions as specified, whereas the present manuscript revision uses existing artifacts only.

D Canonical T1 References and Information Interventions

Training and evaluation contract. The official T1 split removes UPFD from both training and validation, giving 17,728/2,133/2,150 CN train/validation/test records. UPFD's unavailable mechanism labels must not be interpreted as negative supervision. All canonical RoBERTa numbers below use CN-only training. Historical pooled-training results in Appendix G.1 use the same CN test pool but a different training contract.

Input	Macro-F1 (mean $\pm$ SD)	Δ (points)	95% paired-t CI
Full input	0.5340 ± 0.0024	—	—
Claim only	0.4069 ± 0.0049	-12.71	$[-13.48, -11.94]$
Note only	0.5345 ± 0.0026	$+0.04$	$[-0.35, 0.44]$
Random note	0.2464 ± 0.0035	-28.76	$[-29.48, -28.05]$
Same-mechanism wrong note	0.4507 ± 0.0027	-8.33	$[-8.73, -7.93]$
Trigger-masked note	0.5186 ± 0.0057	-1.54	$[-2.23, -0.86]$

Table 10: Canonical CN-only RoBERTa information audit. Five-seed means and sample SDs; differences use unrounded scores and paired-t intervals with four degrees of freedom.

Input	7	13	42	43	44
Full input	0.5350	0.5343	0.5309	0.5328	0.5372
Claim only	0.3993	0.4106	0.4072	0.4118	0.4057
Note only	0.5374	0.5373	0.5326	0.5328	0.5322
Random note	0.2440	0.2476	0.2494	0.2495	0.2416
Same-mechanism wrong note	0.4546	0.4521	0.4500	0.4477	0.4491
Trigger-masked note	0.5173	0.5220	0.5216	0.5092	0.5230

Table 11: Canonical CN-only RoBERTa macro-F1 by training seed. Corruption conditions use the corresponding full-input checkpoint.

Construction and uncertainty. Claim-only, note-only, and full RoBERTa inputs have separately trained checkpoints. Corruption conditions reuse the full checkpoint. A fixed random donor assignment breaks instance pairing; same-mechanism donors share at least one positive gold label. Records

with no eligible donor retain their original note. The control therefore preserves some correct notes, particularly for all-zero-label records. Trigger masking is case-insensitive over the frozen phrase list. Five-seed intervals use $\bar{d} \pm 2.776 s_d/\sqrt{5}$; they describe seed variability under fixed data.

The four prompted replications use all 2,150 test records, with no T1-specific weight update. Each condition is assessed against that model's own full-input result. Their 95% intervals resample paired records 2,000 times with random seed 0 and recompute macro-F1 differences. Prompt compliance and model adaptation differ from RoBERTa, so the absolute scores are not a controlled comparison of architectures.

Input minus full	Δ (points)	95% paired item CI
Claim only	-17.90	$[-19.62, -16.12]$
Note only	-3.01	$[-4.60, -1.37]$
Random note	-17.77	$[-19.70, -15.88]$
Same-mechanism wrong note	-9.82	$[-11.88, -7.73]$
Trigger-masked note	-1.76	$[-2.62, -0.96]$

Table 12: Qwen3-4B prompted information sensitivity; paired item-bootstrap intervals.

Input minus full	Δ (points)	95% paired item CI
Claim only	-21.28	$[-23.44, -18.79]$
Note only	-3.57	$[-5.13, -1.97]$
Random note	-19.81	$[-21.93, -17.45]$
Same-mechanism wrong note	-12.04	$[-14.39, -9.59]$
Trigger-masked note	-2.01	$[-3.33, -0.70]$

Table 13: Qwen3-32B FP8 prompted information sensitivity, using the completed predictions after retrying eight failed calls. The note-only difference is -3.57 points.

Input minus full	Δ (points)	95% paired item CI
Claim only	-18.02	$[-20.25, -15.61]$
Note only	-3.01	$[-4.65, -1.49]$
Random note	-22.78	$[-25.13, -20.33]$
Same-mechanism wrong note	-16.34	$[-18.51, -14.01]$
Trigger-masked note	-1.24	$[-2.40, -0.16]$

Table 14: Gemini-3.5-Flash prompted information sensitivity.

Input minus full	Δ (points)	95% paired item CI
Claim only	-18.65	$[-21.03, -16.35]$
Note only	-1.98	$[-3.72, -0.34]$
Random note	-22.77	$[-25.07, -20.49]$
Same-mechanism wrong note	-14.78	$[-17.03, -12.37]$
Trigger-masked note	-0.75	$[-1.80, 0.29]$

Table 15: Gemini-3.1-Pro prompted information sensitivity. The trigger-masking interval includes zero.

Model	Full	Claim	Note	Random	Same mech.	Masked
Qwen3-4B	0.3999	0.2209	0.3698	0.2222	0.3018	0.3823
Qwen3-32B	0.4237	0.2109	0.3880	0.2256	0.3033	0.4036
Gemini Flash	0.4558	0.2755	0.4257	0.2280	0.2924	0.4433
Gemini Pro	0.4452	0.2586	0.4254	0.2175	0.2974	0.4376

Table 16: Absolute macro-F1 for the four prompted replications on all 2,150 CN test records. These are not adaptation-controlled comparisons with RoBERTa.

Modern full/public reference. The Qwen3-4B QLoRA reference is a six-label classifier, distinct from the zero-shot prompted replication and from the binary source model. It uses a trainable classification head, sigmoid probabilities, threshold 0.5, and class-weighted binary cross-entropy with training-derived positive weights capped at 20. LoRA has rank 16, scale 32, and dropout 0.05 on attention and MLP projections, with NF4 double quantization and bfloat16 computation. Training uses two epochs, learning rate 10^{-4}, effective batch 32 (8 with accumulation 4), 512 tokens, and 5% warmup followed by linear decay. The best validation macro-F1 checkpoint is selected from twice-per-epoch evaluation.

Training	View	7	13	42	Mean ± SD
CN-only	full	0.5577	0.5583	0.5510	0.5557 ± 0.0040
CN-only	tier1	0.5537	0.5580	0.5507	0.5542 ± 0.0037
Historical pooled	full	0.5486	0.5525	0.5599	0.5537 ± 0.0057
Historical pooled	tier1	0.5568	0.5479	0.5566	0.5537 ± 0.0051

Table 17: Qwen3-4B QLoRA references on the CN-2150 test pool. CN-only training is canonical; pooled runs are historical robustness comparisons. Full and Tier-1 models use their corresponding input representations.

RoBERTa's full-view references are 0.5340 ± 0.0024 under CN-only training and 0.5279 ± 0.0023 under historical pooled training. Its historical public pooled result is 0.5291 ± 0.0031; CN-only public RoBERTa was never run. The Qwen full/public scores do not license substituting one view's reference for another model or task.

D.1 External Evidence Regimes

The diagnostic uses processed AVeriTeC, FEVER, and PubHealth records with nonempty claim and evidence. A proportional label-stratified sample targets 1,500 records per benchmark. Evidence is capped at 220 words. Targets use each benchmark's available verdict classes, not T1's six mechanism labels. Each model receives the same six input conditions within a benchmark; the same-label donor uses verdict labels rather than mechanisms. Trigger masking includes both support and refutation phrases.

Benchmark	n	Full	Claim	Evidence	Random	Same label	Masked
averitec	1500	0.5506	0.3338	0.3473	0.2015	0.2204	0.5412
fever	1500	0.7530	0.5337	0.3076	0.2825	0.3042	0.7530
pubhealth	1500	0.4377	0.2465	0.3360	0.1408	0.1514	0.4388

Table 18: Qwen3-4B external macro-F1. Same-label input supplies another record's evidence with the same verdict label. Full means claim plus evidence.

Benchmark	n	Full	Claim	Evidence	Random	Same label	Masked
averitec	1500	0.6156	0.3425	0.3619	0.0916	0.0859	0.5971
fever	1500	0.8020	0.5074	0.2444	0.2467	0.2614	0.8020
pubhealth	1500	0.4199	0.3255	0.3754	0.0400	0.0498	0.4179

Table 19: Qwen3-32B FP8 external macro-F1 under the same condition definitions.

Model	Benchmark	Evidence-only Δ	95% paired CI
qwen3-32b	averitec	-25.37	$[-28.84, -21.91]$
qwen3-32b	fever	-55.75	$[-58.27, -53.22]$
qwen3-32b	pubhealth	-4.45	$[-6.80, -2.15]$
qwen3-32b	VeriBench-Wild	-3.57	$[-5.13, -1.97]$
qwen3-4b	averitec	-20.34	$[-23.75, -17.15]$
qwen3-4b	fever	-44.54	$[-47.40, -41.48]$
qwen3-4b	pubhealth	-10.17	$[-12.72, -7.67]$
qwen3-4b	VeriBench-Wild	-3.01	$[-4.60, -1.37]$

Table 20: Evidence-only changes from full input; 95% paired item-bootstrap intervals. VERIBENCH-WILD uses its full CN test pool and mechanism labels; external verdict tasks use their own sampled pools.

The external scorer reads the first recognized valid verdict. Refusals and unparsed answers receive the sampled task's majority label, preserving the denominator; the stored files retain these flags. Intervals use 2,000 paired item resamples with seed 0. The contrast is illustrative: label spaces, evidence selection, and truncation differ across datasets, so it does not identify a universal effect of evidence type.

E SAME-BACKBONE GENERALIZATION AND ADAPTATION

Pools and scoring. All conditions start from `Qwen/Qwen3-4B-Instruct-2507`. Unmatched training contains 4,078 UPFD and 17,284 CN records; evaluation uses 1,003 and 4,189. Strict matching retains 1,469 and 1,127 training records and 333 and 273 evaluation records. Evaluation is the validation-plus-test union inherited from the source audit, not the official T1 test. No evaluation labels select the QLoRA checkpoint. Matching uses observed covariates and retains original split assignments; Appendix G.4 gives the search and balance diagnostics.

Each claim is capped at 200 tokens with the same tokenizer. Both adaptation regimes choose between `SUPPORTED` and `REFUTED` by summed continuation log probability, including every token of each candidate. A tie selects `REFUTED`. The runner uses left padding with explicit position IDs. It checks batched against individual scoring; no response parser enters the binary metric.

Pool	Regime	$U \to U$	$U \to C$	$C \to C$	$C \to U$	G	95% CI
Unmatched	ICL	0.6529	0.5006	0.5339	0.6248	3.07	$[-5.51, 11.66]$
Unmatched	QLoRA	0.8058	0.4531	0.7274	0.3690	35.56	$[31.67, 39.45]$
Strict	ICL	0.6119	0.4986	0.5096	0.5743	2.43	$[-1.20, 6.06]$
Strict	QLoRA	0.7392	0.5351	0.5466	0.4741	13.83	$[8.21, 19.45]$

Table 21: Same-backbone binary macro-F1 and home-source gap G in points. ICL means average demonstration seeds 7/13/42; QLoRA means average training seeds 7/13/42. Gap intervals use $t_{0.975,2} = 4.303$.

Pool	Regime	7	13	42
Unmatched	ICL	1.287	7.056	0.874
Unmatched	QLoRA	33.906	35.744	37.021
Strict	ICL	3.635	2.847	0.804
Strict	QLoRA	13.535	11.735	16.230

Table 22: Home-source gap in points for each demonstration/training seed.

Pool	Precision	UPFD	CN
Unmatched	bfloat16	0.5869	0.4570
Unmatched	NF4	0.5525	0.4628
Strict	bfloat16	0.5675	0.4467
Strict	NF4	0.5392	0.4277

Table 23: Deterministic zero-shot binary macro-F1. With no source-conditioned adaptation there is no bidirectional home-source G; UPFD–CN accuracy differences are contextual difficulty controls.

Adaptation and precision. ICL uses 16 training demonstrations, balanced as eight supported and eight refuted, fixed by seed and reused for both target sources. ICL inference uses bfloat16 weights. QLoRA uses NF4 double quantization and bfloat16 computation, with rank 16, scale 32, dropout 0.05, and adapters on query/key/value/output and gate/up/down projections. Training uses two epochs, learning rate 10^{-4}, micro-batch 16, accumulation 2, zero weight decay, 5% warmup, and linear decay. Only the assistant label continuation and end token contribute to training loss. The final checkpoint is evaluated.

The zero-shot NF4 control makes precision effects visible but does not equalize all implementation and supervision differences between ICL and QLoRA. The comparison concerns adaptation regime. With three seeds, the t intervals are necessarily imprecise; per-seed item-bootstrap intervals are stored separately and are not substituted for seed uncertainty.

Pool	Seed	n	Gold SUPPORTED (%)	Predicted (%)
Unmatched	7	1003	55.7	9.0
Unmatched	13	1003	55.7	7.9
Unmatched	42	1003	55.7	5.4
Strict	7	333	56.2	54.1
Strict	13	333	56.2	51.7
Strict	42	333	56.2	32.1

Table 24: CN-trained QLoRA evaluated on UPFD: gold and predicted class prevalence. Matching changes the prediction distribution as well as the pool; these rates do not identify a learned prior.

On unmatched UPFD, predicted supported prevalence falls to 5.4–9.0%, against 55.7% gold. On strict matched UPFD it ranges from 32.1% to 54.1%, against 56.2% gold. This diagnostic is consistent with source-conditioned decision behavior. It neither locates a mechanism inside the model nor assigns the transfer deficit solely to training.

F Human-Grounded Scoring and Parser Sensitivity

Human workflow. Three independent raters each labeled all 476 mechanism judgments from the same 384 generated items. Each used the same annotation page, batch, target definitions, and rules, with an independently randomized order. The interface hid generator and automatic-judge identities. Raters saw the source claim, generated text, target mechanism, and its definition. They worked independently and selected YES, NO, NEUTRAL, or UNCERTAIN. YES requires the target mechanism; NEUTRAL records essentially unchanged meaning, and UNCERTAIN permits unresolved judgment. Rater identities are anonymized as R1–R3.

The six definitions are those of the generation evaluator in Appendix H, not the distinct original typology-label study. Binary human-majority YES requires at least two YES votes. NO, NEUTRAL, and UNCERTAIN count as not-YES for this reduction, without being treated as interchangeable in four-label agreement.

Rater	Overall YES (%)	Qwen gen. (%)	Mistral gen. (%)
R1	18.3	15.7	20.8
R2	15.3	12.7	17.9
R3	23.1	17.8	28.3

Table 25: Per-rater YES rates on all 476 judgments. Rater IDs are anonymized.

Rater pair	Binary Cohen κ
R1/R2	0.640
R1/R3	0.432
R2/R3	0.471

Table 26: Pairwise Cohen agreement after reducing each rater's response to YES versus not-YES.

Mechanism	n	Four-label κ	Binary κ
Factual error	93	0.495	0.383
Manipulated media	69	0.349	0.328
Outdated information	78	0.085	0.034
Missing context	89	0.228	0.238
Unverified claim	85	0.712	0.726
Satire	62	0.455	0.565

Table 27: Human Fleiss agreement by mechanism. The 0.03–0.73 range in the main text is binary agreement; four-label agreement differs. Counts include multiple target mechanisms for some generated items.

Across all judgments, Fleiss κ is 0.534 for four labels and 0.507 for the binary reduction. Nominal Krippendorff α rounds to the same values, 0.534 and 0.507. Human-majority acceptance is 78/476 (16.4%): 46/240 Mistral and 32/236 Qwen judgments. Their 5.61-point difference has 95% interval $[-0.65, 12.21]$. The frozen bootstrap uses 5,000 item-clustered resamples, stratified by generator, using seed 0; all mechanism judgments for a sampled item remain together. The binary Fleiss interval is $[0.424, 0.582]$. Shared source records can induce dependencies beyond these item clusters.

Judge (strict)	YES (%)	κ	Bal. acc.	Precision	Recall	F1
Mistral-7B	63.4	0.103	0.611	0.212	0.821	0.337
Qwen2.5-0.5B	45.0	0.117	0.599	0.224	0.615	0.329
Phi-3-medium	30.5	0.367	0.747	0.386	0.718	0.502
Qwen3-4B	0.4	0.017	0.505	0.500	0.013	0.025
Qwen3-32B	20.4	0.448	0.746	0.495	0.615	0.549
Gemini Flash	15.1	0.383	0.686	0.500	0.462	0.480
Gemini Pro	16.8	0.469	0.737	0.550	0.564	0.557
Grok-4.7	18.3	0.509	0.766	0.563	0.628	0.594

Table 28: Eight judges against binary human majority under strict parsing, all 476 judgments. Balanced accuracy averages sensitivity and specificity; precision/recall/F1 concern YES. Table order is fixed rather than ranked by quality.

Agreement and acceptance are different measurements. For example, the Mistral judge's 4.0-point generator difference is near the human 5.6-point difference, yet its item-level κ is only 0.103. Gemini Pro's aggregate acceptance is near the human rate, but this does not imply identity of accepted items. Human disagreement further limits interpretation of the majority as a correctness target.

Strict and lenient scoring. The strict rule searches case-insensitively for `VERDICT: YES|NO|NEUTRAL` and maps unmatched responses to NO. The lenient rule additionally recovers a verdict on the first nonempty line, optionally bolded or punctuated, when no explicit verdict field is found. Unrecoverable responses remain not-YES in this paper's fixed-denominator rates. Lenient parsing recovers all 476 responses for each new judge in this run.

Judge	Strict YES	Lenient YES	Template	Qwen gen.	Mistral gen.
Mistral-7B	63.4	—	—	61.4	65.4
Qwen2.5-0.5B	45.0	—	—	38.1	51.7
Phi-3-medium	30.5	—	—	23.7	37.1
Qwen3-4B	0.4	31.9	1.1	23.3	40.4
Qwen3-32B	20.4	20.4	100.0	15.7	25.0
Gemini Flash	15.1	15.3	99.8	11.0	19.6
Gemini Pro	16.8	16.8	100.0	11.4	22.1
Grok-4.7	18.3	18.3	100.0	13.1	23.3

Table 29: All entries are percentages. Template is the stored response-template compliance flag, available only for the five new judges. Generator columns use lenient verdicts for new judges and stored strict verdicts for the earlier three. Every overall denominator is 476; generator denominators are 236 Qwen and 240 Mistral. Dashes mean unavailable, not zero.

Qwen3-4B has 5/476 template-compliant responses, with strict YES 2/476 and lenient YES 152/476. Its lenient alignment is $\kappa = 0.312$, balanced accuracy 0.715, YES precision 0.349, recall 0.679, and F1 0.461. Gemini Flash changes from 72 to 73 accepted judgments; its lenient alignment is 0.378/0.684/0.493/0.462/0.477 in the same metric order. Qwen3-32B, Gemini Pro, and Grok have identical strict/lenient metrics.

These are paired scoring rules applied to stored responses; neither models nor generations are rerun. The three original judges preserve parsed verdicts rather than full responses, so a comparable lenient reparse is unavailable. Their older stored-field proxy is reported separately in Appendix H.2 and must not be interpreted as measured parser failures.

G Reference Results and Audit Details

Historical architecture comparison. Table 30 lists the five stable pooled T1 reference configurations. The range is $100(0.5327355403 - 0.5231393035) = 0.9596$ percentage points. These include two RoBERTa runs, so they should not be described as five distinct model architectures. The comparison is descriptive and excludes unstable or collapsed runs. It is not a statistical equivalence test. The historical pooled-test RoBERTa mean is 0.5327. The same pooled-trained model scores 0.5279 on the CN-only test pool; the official CN-only-trained reference is 0.5340.

Task / pool	Reference	Mean ± SD	Stored seeds
T1 pooled	RoBERTa	0.5327 ± 0.0021	5
T1 pooled	Legacy RoBERTa	0.5315 ± 0.0023	5
T1 pooled	DeBERTa-base	0.5299 ± 0.0018	5
T1 pooled	ModernBERT	0.5234 ± 0.0066	5
T1 pooled	Multilingual DeBERTa	0.5231 ± 0.0035	5
T-Prop	GraphGPS	0.9683 ± 0.0015	5
T-Prop	TF-IDF logistic regression	0.8051	5 identical
T-Author	History transformer	0.8327 ± 0.0014	5
T-Author	GRU history	0.8323 ± 0.0048	5
T-Author	Mamba–Hawkes history	0.8291 ± 0.0009	5
T-Author	History-statistic logistic regression	0.8152	5 identical
T-Author	Author-rate rule	0.8048	5 identical

Table 30: Selected frozen reference baselines. T1 and T-Prop use macro-F1; T-Author uses ROC-AUC. The T-Author architecture range is about 0.36 points. Deterministic identical outputs are not independent training runs.

G.1 Historical Pooled-Training Information Audit

This subsection preserves the earlier pooled-training controls, not the canonical CN-only training results. The fixed CN-only evaluation pool is defined by excluding the legacy tier. All its records have nonempty note text. The training conditions use their corresponding saved predictions; the corrupted-note conditions evaluate the full-input checkpoints with fixed donor assignments. The

same-mechanism donor shares at least one positive label, not necessarily the full label vector. When no donor is available, the original note remains. In particular, all-zero-label records remain unchanged in that control.

Mechanism	Note-only F1	Full-input F1	Difference (points)
Factual error	0.6443	0.6391	$+0.515$
Manipulated media	0.8229	0.8110	$+1.191$
Outdated information	0.3612	0.3548	$+0.642$
Missing important context	0.6472	0.6398	$+0.734$
Unverified claim as fact	0.4760	0.4722	$+0.374$
Satire	0.2521	0.2507	$+0.139$

Table 31: Historical pooled-training five-seed per-mechanism means on the CN-only test pool. Differences use unrounded scores. Note-only is nominally higher on all six axes; the satire margin is effectively a tie. No axis reverses, but these are not separate claims of significant per-axis superiority.

Input condition	Item-bootstrap 95% CI, seed 42 macro-F1
Claim only	$[0.3838, 0.4176]$
Note only	$[0.5168, 0.5456]$
Claim + correct note	$[0.5170, 0.5461]$
Claim + random note	$[0.2435, 0.2728]$
Claim + same-mechanism wrong note	$[0.4377, 0.4667]$
Claim + trigger-masked note	$[0.5005, 0.5300]$

Table 32: Uncertainty over test records for a fixed checkpoint (10,000 item resamples). These intervals differ from the paired-t seed-difference intervals in Appendix G.2; they do not provide joint seed-and-item uncertainty.

G.2 Historical Pooled-Training Statistical Sensitivity

These are the older pooled-trained models evaluated on CN records, not the canonical CN-only training audit. With five seeds, percentile resampling gives a limited view of training variability. We retain the historical bootstrap for continuity, not as the sole basis for sign or significance claims. Paired-t intervals use $\bar{d} \pm t_{0.975,4} s_d/\sqrt{5}$ and assume approximately normal seed differences; that assumption is also hard to assess at this sample size.

Input minus full	7	13	42	43	44	SD
Claim only	-12.750	-12.092	-13.112	-12.499	-12.420	0.382
Note only	$+0.578$	$+0.282$	-0.033	$+0.901$	$+1.268$	0.509
Random note	-26.720	-26.987	-27.358	-26.838	-26.851	0.247
Same-mechanism wrong note	-7.470	-7.869	-7.923	-7.615	-7.165	0.309
Trigger-masked note	-0.818	-0.958	-1.653	-0.617	-1.529	0.453

Table 33: Raw information-audit differences in percentage points, by seed. The note-only comparison has one negative difference.

Input minus full	Mean (points)	Five-point bootstrap CI	Paired-t CI
Claim only	-12.575	$[-12.895, -12.289]$	$[-13.049, -12.101]$
Note only	$+0.599$	$[+0.215, +0.992]$	$[-0.033, +1.232]$
Random note	-26.951	$[-27.156, -26.794]$	$[-27.257, -26.645]$
Same-mechanism wrong note	-7.609	$[-7.840, -7.367]$	$[-7.993, -7.224]$
Trigger-masked note	-1.115	$[-1.464, -0.778]$	$[-1.678, -0.552]$

Table 34: 95% intervals over five seed differences. The percentile bootstrap uses 10,000 resamples. Interval-method sensitivity matters for note-only: its small positive average is not established superiority.

The four evidence-removal/corruption comparisons have the same sign in all five seeds and under both interval methods. For note-only, the mean is $+0.599$ points with SD 0.510, but the paired-t

interval includes zero. The paper therefore emphasizes the large losses from absent or mismatched evidence.

G.3 Historical Pooled-Training Truncation Control

At the 256-token limit, 20.3% of full test inputs truncate, compared with 6.9% of note-only inputs and 0.8% of claim-only inputs. Training truncation rates are 21.2%, 7.3%, and 0.8%, respectively. Overflow is removed from the end, so prepending the claim can discard note content. Among affected full-input test records, 67.2% of note tokens remain on average.

The frozen control rescores stored predictions on 1,714 records where both full and note-only inputs fit. It uses the same five model seeds without retraining.

Comparison	Full-pool delta (points)	Truncation-free delta (points)
Note-only minus full	$+0.599$ $[+0.215, +0.992]$	$+0.490$ $[+0.068, +0.911]$
Claim-only minus full	-12.575 $[-12.895, -12.289]$	-14.045 $[-14.480, -13.663]$

Table 35: Truncation control; 95% intervals use 10,000 paired-seed bootstrap resamples. The small note-only mean difference remains, while the claim-only deficit remains large. Five-point bootstrap intervals are retained for continuity, not as a sole sign test. The subset differs in composition, so subtracting these estimates does not identify a causal contribution of truncation.

The note-only bootstrap interval is marginally positive, but that method is sign-sensitive at five seeds. We interpret the truncation control as preserving similar average note-only/full performance and a large claim-only deficit, not establishing note-only superiority.

G.4 Matching Details and Supporting RoBERTa Results

Coarsened exact matching uses language, binary verdict, length bins, and 30 topic clusters from frozen MiniLM embeddings. Combined matching also rebalances the marginal verdict distribution. The strict variant uses 120 log-length bins and retains the largest candidate pool with absolute length standardized mean difference below 0.25. Candidate selection uses covariate balance, not held-out model scores. Existing split membership is preserved after selecting records.

Setting	UPFD records	CN records	Length SMD	Gap, points [95% CI]
Harmonized unmatched	5,081	21,473	-1.468	31.01 $[29.21, 32.08]$
English-only	5,081	16,411	-1.581	29.65 $[28.23, 31.49]$
Verdict-balanced	4,572	15,980	-1.549	29.95 $[27.38, 32.09]$
Length-matched	5,081	16,198	-1.120	32.80 $[31.10, 34.59]$
Topic-matched	5,081	17,418	-1.439	30.72 $[28.97, 32.61]$
Combined	3,364	3,362	-0.496	21.61 $[19.35, 23.88]$
Strict combined	1,802	1,400	-0.227	19.03 $[14.70, 22.31]$

Table 36: Matched-pool sizes across all splits, length standardized mean difference (SMD), and five-seed bidirectional gaps. Intervals are continuity summaries from 10,000 resamples of five seed values, not precise population intervals. The unmatched row uses the harmonized control. Single-covariate restrictions leave large gaps; the strict pool is still only partially balanced.

Candidate	UPFD records	CN records	Length SMD
10 log-length bins	3,368	3,318	−0.473
20 bins	2,868	2,912	−0.426
40 bins	2,422	2,238	−0.366
60 bins	2,226	1,894	−0.287
80 bins	2,006	1,662	−0.264
120 bins (selected)	1,802	1,400	−0.227
40 bins + support trim 0.3	846	930	−0.090
40 bins + support trim 0.5	490	514	−0.018
60 bins + support trim 0.3	854	888	−0.031

Table 37: Frozen matching search. All candidates have zero marginal support-rate SMD. Tighter balance is possible at the cost of substantially fewer retained records; these alternative candidates are balance diagnostics, not additional trained transfer results.

Strict matched training uses 1,469 UPFD and 1,127 CN records. The held-out records also change. Neither training-data reduction nor evaluation-set difficulty is isolated from changes in source-specific signal. The source-classifier diagnostic holds its covariates fixed to language, character length, and verdict across unmatched, combined, and strict pools. Its AUROCs are 0.888, 0.627, and 0.561, respectively. Topic is not included in this diagnostic, even though topic participates in matching.

G.5 Supporting RoBERTa Source Control and Protocol Sensitivity

The harmonized unmatched run reproduces the historical record selection while using the matched runner verbatim for model training and evaluation. It retains 4,078 UPFD and 17,284 CN training records; the evaluation pools contain 1,003 and 4,189 records. Only the pool source and output destination change relative to the matched runner. This removes the implementation mismatch within the RoBERTa comparison; the primary source/adaptation experiment now uses Qwen (Appendix E).

Cell	Historical unmatched	Harmonized unmatched	Strict matched
UPFD → UPFD	0.8102 ± 0.0097	0.8051 ± 0.0153	0.7467 ± 0.0243
UPFD → CN	0.4606 ± 0.0239	0.4870 ± 0.0101	0.4964 ± 0.0331
CN → CN	0.7256 ± 0.0059	0.7181 ± 0.0099	0.5705 ± 0.0470
CN → UPFD	0.4174 ± 0.0156	0.4159 ± 0.0362	0.4402 ± 0.0607

Table 38: Macro-F1, five-seed mean ± SD. Historical scores are protocol sensitivity only; the harmonized column is the unmatched reference for this RoBERTa comparison.

Gap / paired change	7	13	42	43	44
Historical gap	33.879	34.645	30.116	33.631	32.161
Harmonized gap	31.661	32.353	31.893	27.498	31.649
Strict gap	22.414	10.884	18.013	20.988	22.855
Harmonized − historical	−2.218	−2.291	1.778	−6.133	−0.512
Strict − harmonized	−9.248	−21.469	−13.880	−6.510	−8.793

Table 39: Raw source-audit seed values in percentage points. Every seed has a large remaining strict gap and a smaller gap after matching.

Gap / paired change	Mean ± SD	Five-point bootstrap CI	Paired-t CI
Historical gap	32.886 ± 1.791	$[+31.343, +34.136]$	$[+30.662, +35.111]$
Harmonized gap	31.011 ± 1.984	$[+29.207, +32.077]$	$[+28.547, +33.475]$
Strict gap	19.031 ± 4.932	$[+14.704, +22.305]$	$[+12.907, +25.155]$
Harmonized − historical	-1.875 ± 2.899	$[-4.226, +0.179]$	$[-5.475, +1.724]$
Strict − harmonized	-11.980 ± 5.941	$[-16.959, -7.971]$	$[-19.357, -4.603]$

Table 40: Source-audit intervals in points. Bootstrap intervals retain the 10,000-resample, five-seed continuity analysis; paired-t intervals expose the sensitivity of their width.

The historical unmatched gap is 32.89 points, versus 31.01 under the common runner. Their mean difference is -1.88 points, with intervals including zero. This comparison changes implementation details, not record selection; it does not isolate the scheduler from every other code-path difference. The RoBERTa strict-minus-harmonized change is -11.98 points. Reduced training size and changed evaluation records remain inseparable from altered source-specific signals. Most narrowing comes from lower within-source scores, not substantially improved transfer.

G.6 Original Three-Judge Interaction and Sampling

The sample contains 28 single-axis items for each generator–axis cell and 24 multi-axis items per generator. Exact duplicate generated text is removed within generator before sampling. Degenerate/spam outputs are retained when selected; 85 such items occur in the frozen 384-item sample. The sampling seed and selected-item manifest accompany the results. Items can be correlated through shared source records or checkpoints, beyond the within-item axis clustering addressed here.

Let $p_{j,g}$ be the YES-rate for judge j on generator family g. With M for Mistral and Q for Qwen, the raw interaction is

$$(p_{M,M} - p_{M,Q}) - (p_{Q,M} - p_{Q,Q}) = -0.09555.$$

Its 95% item-bootstrap interval is $[-0.229, +0.041]$, from 10,000 resamples that retain all axis judgments belonging to each sampled item. Phi-3 is excluded from the interaction analysis because neither generator shares its family.

The logistic specification is

$$\operatorname{logit} \Pr(\text{YES}) = \alpha + \beta_{\text{judge}} + \gamma_{\text{generator}} + \eta_{\text{axis}} + \theta\, \mathbf{1}\{\text{family match}\}.$$

The model uses 952 judgments from the two crossed judges. The family-match coefficient is $\theta = -0.199$ (log-odds units), with item-clustered bootstrap interval $[-0.509, +0.084]$ from 500 successful resamples. This is not a percentage-point coefficient. Both intervals include zero. This original interaction estimates propensity rather than correctness. Human annotations now cover this entire sample; Appendix F reports the separate human-alignment analysis.

G.7 Provenance Labels and Strict Evaluation

The strict test pool contains 436 `L_HIGH` positives and 748 `NEG` negatives. Average precision is computed on this identical pool for all reported systems. Its positive prevalence is 0.3682; the empirical random baseline differs slightly because it averages finite random rankings.

System	Strict AUPRC, mean ± SD	Stored seeds
Desklib	0.8483 ± 0.0123	5
RoBERTa	0.8392 ± 0.0086	5
TF-IDF logistic regression	0.7782	5 identical
Length logistic regression	0.6184	5 identical
Binoculars	0.3828	1
Random	0.3737 ± 0.0097	5
Fast-DetectGPT	0.3733	1
Constant score	0.3682	5 identical

Table 41: Strict-protocol provenance reference scores. AUPRC means average precision here. Single-seed zero-shot outputs are not presented as five-seed estimates.

The frozen construct rule combines the gold manipulated-media label with media keywords in claim or note text. Among positives it assigns 299 records (68.6%) to media-referencing, 54 (12.4%) to text-native, and 83 (19.0%) to mixed/unclear. Across the full pool, textual media cues occur in claim only for 20 records, note only for 396, both for 48, and neither for 720. Thus, the construct rule relies heavily on information forbidden to the classifier. It describes the positive labels, not the features the claim-only model detects.

Strict claim-text masking sensitivity. On the same strict pool, stored masked RoBERTa predictions yield mean AUPRC 0.8130 versus 0.8392 unmasked. The mean change is -2.62 points; the historical

five-seed percentile interval is $[-3.65, -1.47]$ (10,000 resamples). Masked seed scores, ordered 7/13/42/43/44, are 0.8002/0.7994/0.8165/0.8240/0.8248. Every seed declines, while the masked mean remains above the 0.6184 length reference. Explicit provenance phrases account for some signal, but not all. Both conditions use claim text only; this test does not permit evidence-note input.

G.8 Supporting Diagnostics

Length probes. The length-longer heuristic scores 0.521 on 1,000 LIAR pairs, with interval $[0.490, 0.552]$, and 0.508 on 1,000 AVeriTeC pairs, with interval $[0.477, 0.539]$. These frozen intervals use the normal approximation $p \pm 1.96\sqrt{p(1-p)/1000}$, not a bootstrap. Pair draws can reuse claims, so the intervals do not model claim-level dependence. The heuristic is near chance on these two probes; no general conclusion about shortcuts disappearing with sample size follows.

Propagation reference contrast. GraphGPS scores 0.9683 ± 0.0015 macro-F1, compared with 0.8051 for TF-IDF logistic regression on the legacy verdict task. The descriptive margin is about 16.3 points. The text reference uses a different input view from the graph model, so the margin is not an isolated architecture effect.

Cascade statistics. On the UPFD legacy source subsets, Cliff's δ for supported-minus-refuted cascade size is $+0.6087$ in GossipCop ($n = 4,871$) and -0.1517 in PolitiFact ($n = 284$). This is a source-conditional cascade statistic, not a reversal of classifier F1. It does not describe Community Notes propagation.

Language slices. The historical pooled-training T1 language analysis finds Spanish, Portuguese, and Japanese above English. This supports the task-specific observation that English is not the easiest T1 slice. The same statement does not hold for T2-HP. Those language results use their own task pools and should not be mixed with the strict-protocol provenance ladder. Small slices and differing label composition limit comparisons.

G.9 Per-Seed Reference Scores

The T1 values below are historical pooled-training references; CN in a row name identifies evaluation records only. The following tables report the stored seed-level scores behind the reference summaries. Pooled and Community-Notes-only results use different evaluation records and must remain separate comparisons.

Task	Reference	7	13	42	43	44
T1 pooled	RoBERTa	0.5330	0.5321	0.5366	0.5312	0.5308
T1 pooled	Legacy RoBERTa	0.5302	0.5280	0.5336	0.5343	0.5313
T1 pooled	DeBERTa-base	0.5318	0.5272	0.5285	0.5315	0.5302
T1 pooled	ModernBERT	0.5168	0.5275	0.5147	0.5321	0.5259
T1 pooled	Multilingual DeBERTa	0.5244	0.5222	0.5260	0.5263	0.5167
T-Prop	GraphGPS	0.9678	0.9699	0.9659	0.9679	0.9699
T-Author	Transformer	0.8317	0.8315	0.8354	0.8328	0.8320
T-Author	GRU	0.8332	0.8364	0.8359	0.8328	0.8232
T-Author	Mamba–Hawkes	0.8286	0.8292	0.8283	0.8288	0.8308

Table 42: Stored per-seed reference scores. T1 is pooled macro-F1; T-Prop is binary macro-F1; T-Author is ROC-AUC. These values are transcribed from the corresponding frozen summary artifacts without rerunning models.

Condition	7	13	42	43	44
T1 CN claim-only	0.4003	0.4062	0.4007	0.4011	0.4026
T1 CN note-only	0.5336	0.5299	0.5315	0.5351	0.5395
T1 CN full	0.5278	0.5271	0.5319	0.5261	0.5268
T2-HP RoBERTa	0.8351	0.8404	0.8343	0.8309	0.8554
T2-HP Desklib	0.8276	0.8418	0.8525	0.8605	0.8592

Table 43: Historical pooled-trained, CN-evaluated information-channel references (macro-F1) and strict provenance references (AUPRC). Source-specific values must not be interchanged with the pooled T1 table.

H Generation and Evaluator Prompts

Generator input. The Mistral training/generation script supplies the following template. The placeholders are the evidence note and a comma-separated list of target mechanisms. The original claim is a training target where applicable, not part of the generator prompt.

```
<s>[INST] Rewrite the following true claim/explanation into a misleading
↪  claim that exhibits axis(es): {axes}.
Note: {note}
Misleading claim: [/INST]
```

The instruction's description of an explanation as true is part of the frozen prompt, not an independent validation of every note. The Qwen system uses the same note-and-mechanism task contract with its own chat formatting.

Evaluator input. The source claim is retrieved separately as evaluator context. The audit uses the following user-prompt text; local models apply their native chat templates, and hosted models use the recorded provider settings. The template below preserves the words and line breaks produced by the frozen `build_prompt` function; visual line wrapping is performed by LaTeX.

```
You are a careful misinformation-typology annotator. Decide whether a
↪  rewrite of an original claim successfully introduces the
↪  misleadingness type '{short_name}'.

Definition of '{short_name}': {definition}

Original claim:
"""
{source_claim}
"""

Rewrite (intended to exhibit '{short_name}'):
"""
{generated_text}
"""

Answer with EXACTLY one token on a single line:
  YES     -- the rewrite clearly exhibits the target axis
  NO      -- the rewrite does not exhibit the target axis
  NEUTRAL -- the rewrite is essentially identical in meaning to the
   ↪  original (no axis-specific change introduced)

Then on a second line give a 1-sentence justification (<= 25 words).
Format strictly:
VERDICT: <YES|NO|NEUTRAL>
REASON: <one sentence>
```

Short name	Definition substituted verbatim
factualError	the rewrite asserts something that is factually wrong (a verifiably false claim).
manipulatedMedia	the rewrite explicitly references or implies edited, fake, doctored, deepfake, AI-generated, or otherwise manipulated images, video, or audio.
outdated	the rewrite presents information that was once true but is no longer true now (stale stats, expired status, revoked policy, old date framed as current).
missingContext	the rewrite omits crucial context that, if included, would change how a reader interprets the claim (cherry-picked numbers, no time frame, no comparison group, no caveats).
unverified	the rewrite presents speculation, rumor, anonymous sourcing, or an unverified allegation as if it were established fact.
satire	the rewrite is framed as joke, parody, irony, sarcasm, or obvious satire (e.g. Onion-style).

Table 44: Frozen evaluator mechanism definitions. In particular, its satire definition is broader than satire passed off as serious in the annotation rubric. This is part of the evaluator protocol, not a silently harmonized label definition.

Models and decoding. The three original judge identifiers are `mistralai/Mistral-7B-Instruct-v0.3`, `Qwen/Qwen2.5-0.5B-Instruct`, and `microsoft/Phi-3-medium-4k-instruct`. Each uses its native chat template with an added generation prompt, greedy decoding (`do_sample=False`), and at most 60 new tokens. The GPU runner uses bfloat16. It does not impose the older judge runner's character truncation. Prompt content is held fixed, but tokenization and chat formatting remain model-specific.

Additional models and inference settings. The four prompted T1 replications and five additional T0 judges use the saved model aliases `qwen3-4b`, `qwen3-32b-fp8`, `gemini-3.5-flash`, `gemini-3.1-pro`, and, for T0, `grok-4.7`. The Qwen models are `Qwen/Qwen3-4B-Instruct-2507` and `Qwen/Qwen3-32B-FP8`. Provider identifiers are `gemini-3.5-flash`, `gemini-3.1-pro-preview`, and requested `grok-4.7` (returned version `grok-4.7-build`). Grok is evaluated with valid xAI API-key access under the stated setup; the saved configuration disables tools and search and requests low reasoning effort.

Qwen inference disables thinking through the chat-template setting. The shared frontier client requests temperature 0 and 128 output tokens for information and evaluator prompts. Gemini Flash uses minimal thinking; Gemini Pro uses low thinking with a 2,048-token allowance added for reasoning. These provider settings are not identical cognitive budgets. External verdict prompts use a 16-token response budget. The new evaluator files retain strict and lenient verdicts and response excerpts; the original three retain parsed verdicts and reasons.

Parsing and aggregation. A case-insensitive search extracts `VERDICT: YES`, `VERDICT: NO`, or `VERDICT: NEUTRAL`. Unparsed outputs default to NO. The overall YES-rate pools axis judgments, so an item with several target axes contributes several judgments. The original family-match and new human-comparison bootstraps keep judgments clustered by item. This metric differs from an equal-weight mean over the six mechanisms and from a per-item mean of within-item axis rates.

The evaluator prompt describes outputs as rewrites even though the generator receives note text rather than an original claim. Reproduction preserves this asymmetry. A future evaluator that removes the original claim or changes the mechanism definitions should be reported as a different protocol.

H.1 Information and Binary-Transfer Prompts

The prompted T1 audit uses the following template. The definition block contains the six definitions above, in the same order, with each occurrence of *the rewrite* replaced by *the item* and each line formatted as `- shortName: definition`. The body contains the permitted claim and/or the condition-specific note; omitted channels are absent rather than empty.

```
You are a careful misinformation-typology annotator working from the
↪ X/Community Notes typology. Decide, for each of the six
↪ misleadingness types below, whether it applies to the item shown.

Types:
```

```
{definition_block}

The types are NOT mutually exclusive: answer each one independently, and
↪  answer NO to every type that does not apply.

{body}

Reply with EXACTLY six lines and nothing else, in this order and format:
factualError: <YES|NO>
manipulatedMedia: <YES|NO>
outdated: <YES|NO>
missingContext: <YES|NO>
unverified: <YES|NO>
satire: <YES|NO>
```

Each included body block is formatted as follows, with a blank line between blocks:

```
CLAIM:
"""
{claim}
"""

FACT-CHECK NOTE:
"""
{condition_note}
"""
```

The parser searches case-insensitively for each named label followed by a colon and YES or NO, allowing whitespace and asterisks before the verdict. Missing labels default to NO. Complete six-label parsing ranges from 99.67% to 100% across the completed runs; final API-failure rates are zero.

The binary source audit uses the following user content inside the native chat template. The demonstration block is repeated 16 times for ICL and omitted for zero-shot and QLoRA inference. Candidates are scored as continuations rather than parsed from generated text.

```
You are a fact-checking annotator. Decide whether each social-media
↪  claim is SUPPORTED (substantially accurate as stated) or REFUTED
↪  (misleading, false, or missing context that changes its meaning).
↪  Answer with exactly one word: SUPPORTED or REFUTED.

CLAIM: {demonstration_claim}
VERDICT: {demonstration_label}

CLAIM: {target_claim}
VERDICT:
```

H.2 External Verdict Prompt

The external diagnostic uses the following template. Definitions and the final answer list include only the sampled benchmark's available labels, in sorted order. Its body uses the claim block above and an evidence block headed `EVIDENCE:`, with the same triple-quote delimiters; absent channels are omitted.

```
You are a careful fact-checking annotator. Assign exactly one verdict to
↪  the item below.

Verdicts:
{verdict_definitions}

{body}

Reply with exactly one word on one line, one of: {available_labels}.
```

```
VERDICT:
```

The available definition lines are:

```
- CONFLICTING: the evidence is mixed: parts support and parts contradict
↪  the claim.
- INSUFFICIENT: the evidence does not settle the claim either way.
- REFUTED: the evidence contradicts the claim.
- SUPPORTED: the evidence supports the claim as stated.
```

Frozen masking vocabulary. The T1 mask replaces case-insensitive whole-phrase matches with `[MASK]`, applying longer alternatives first:

```
factually incorrect; not true; false; untrue; fabricated; no evidence;
↪  incorrect; inaccurate; debunked; myth; ai generated; ai-generated;
↪  manipulated; doctored; deepfake; fake image; edited image; altered;
↪  photoshopped; synthetic; not a real; not real; outdated; no longer;
↪  old footage; old video; old image; previously; from an earlier;
↪  years ago; not current; missing context; lacks context; out of
↪  context; without context; important context; context; unverified;
↪  unconfirmed; rumor; alleged; no credible source; no reliable source;
↪  unsubstantiated; satire; satirical; parody; comedic; joke; not meant
↪  to be taken literally
```

The external verdict diagnostic adds the following phrases to the same vocabulary:

```
true; correct; accurate; confirmed; verified; supported; consistent
↪  with; as stated; indeed; does support; corroborated; substantiated;
↪  refuted; contradicted; disproven; no support; does not support;
↪  insufficient evidence; not enough information; cannot be verified;
↪  unclear; conflicting; mixed evidence; partly true; half true;
↪  misleading
```

Historical three-judge response-compliance proxy. Full raw responses were not retained for the three original judges, so their exact parser failures cannot be reconstructed. The five new judges retain strict/lenient verdicts and response records (Appendix F). The frozen scorer saves verdict and reason, mapping an unrecognized verdict to NO. A proxy flags rows whose stored verdict is NO and whose reason is empty. A valid NO without a reason can also trigger this proxy; it is not a measured parser-failure rate.

Judge	Proxy flags / judgments	Frozen YES-rate	Excluding flagged rows
Mistral	0/476	63.4%	63.4% ($n = 476$)
Qwen2.5-0.5B	105/476 (22.1%)	45.0%	57.7% ($n = 371$)
Phi-3	0/476	30.5%	30.5% ($n = 476$)

Table 45: Stored-field compliance proxy, not exact parser classification. Exclusion changes the denominator and is a sensitivity check, not the official evaluator metric.

The Mistral–Phi-3 spread remains 32.98 points under this check because neither judge has flagged rows. Qwen's position within that range is sensitive to the convention. Its proxy flags cover both generator families: 56/240 Mistral and 49/236 Qwen judgments. Excluding flagged rows changes the raw family interaction from -0.096 to -0.153, farther from positive same-family acceptance. No new interval or significance claim is attached to that sensitivity. The historical frozen rates and interaction remain unchanged. Zero proxy flags do not establish perfect formatting or judge correctness.

I ETHICS, ACCESS, AND REPORTING BOUNDARIES

Privacy and access. Public posts remain linkable through text search even after identifiers are hashed. Tier-1 removes direct keys and coarsens timestamps; gated graphs remain separate from public summaries. These measures reduce linkage opportunities without guaranteeing anonymity. Reusers should preserve access restrictions and avoid identifying posters or annotators.

Consent and annotation. Individual poster consent was not obtained and no ethics-board filing was made. Participants in the original label study were research collaborators rather than paid crowdworkers. The new T0 study uses three independent raters. Public posting is not equivalent to individual consent to every downstream research use.

Licenses and scope of redistribution. The release documents CC-BY-SA-4.0 for UPFD-derived material and CC-BY-NC-SA-4.0 for newly derived material. These descriptions do not replace license files or grant blanket rights over third-party posts. Verbatim content and graphs remain subject to their access arrangements.

Potential harms. The corpus can expose readers to offensive, misleading, or sensitive content. Generation models can produce misleading claims beyond their intended evaluation use. Author-history predictions concern dataset attribution labels and should not become judgments about an individual's trustworthiness. Neither the source audit nor the language slices justify claims about demographic groups.

J MEASUREMENT-CONTRACT SCHEMATIC

Figure 5 gives a schematic view of the measurement contract behind a reported score and of the three audits, complementing the overview in Figure 1.

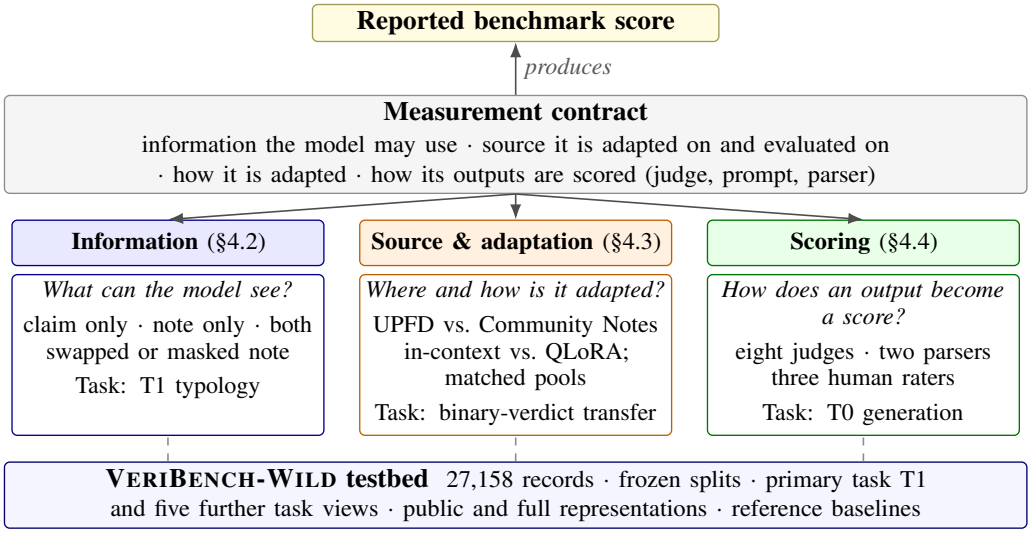


Figure 5: Measurement-contract schematic. A reported score depends on the model and on the settings behind it. We audit three settings on the released data, each on the task where it applies. The audits use different tasks and targets, so they are not additive parts of one error measure.

K RANKING CONSEQUENCES AND PARSER SENSITIVITY ACROSS JUDGES

This appendix asks whether each audited setting changes the *ordering* of systems, not only their scores. The information and generalization analyses use stored predictions only and make no model calls. For the information audit, T1 macro-F1 is recomputed from the stored per-record predictions of the four prompted models on the 2,150 Community Notes test records; every recomputed value matches the stored result. Intervals are percentiles of 5,000 paired item-bootstrap resamples (seed 0), so all models are compared on the same resampled records. For the generalization audit, per-seed cell scores are recomputed from the stored predictions and match the stored four-cell means; within-source and cross-source means are compared with seed-paired t-intervals (three seeds, two degrees of freedom).

Condition	Qwen3-4B	Qwen3-32B	Gem. Flash	Gem. Pro	API − open	τ
Claim + note (full)	40.0	42.4	45.6	44.5	$+3.9$ $[+2.5, +5.2]$	$+1.00$
Claim only	22.1	21.1	27.6	25.9	$+5.1$ $[+3.5, +6.7]$	$+0.67$
Note only	37.0	38.8	42.6	42.5	$+4.7$ $[+3.4, +5.9]$	$+1.00$
Random note	22.2	22.6	22.8	21.7	-0.1 $[-1.7, +1.4]$	$+0.33$
Same-mechanism note	30.2	30.3	29.2	29.7	-0.8 $[-2.2, +0.7]$	-0.67
Masked note	38.2	40.4	44.3	43.8	$+4.8$ $[+3.3, +6.2]$	$+1.00$

Table 46: T1 macro-F1 ($\times 100$) of the four prompted models under each input condition. API − open is the mean Gemini score minus the mean Qwen score, with its paired 95% bootstrap interval. τ is Kendall's τ between the condition's model ordering and the full-input ordering. The API lead holds with correct, missing, or masked evidence and disappears with a wrong note; no single pairwise difference under the two wrong-note conditions is significant.

Pool	Scores	ICL (16-shot)	QLoRA	QLoRA − ICL
Unmatched	within source	59.3	76.7	$+17.3\ [+14.3, +20.3]$
	cross source	56.3	41.1	$-15.2\ [-21.7, -8.6]$
Strict-matched	within source	56.1	64.3	$+8.2\ [+2.7, +13.7]$
	cross source	53.6	50.5	$-3.2\ [-6.7, +0.3]$

Table 47: Mean within-source and cross-source binary-verdict macro-F1 ($\times 100$) for the two adaptation regimes from the same initial Qwen3-4B backbone, with seed-paired 95% t-intervals for the QLoRA − ICL difference. On unmatched pools, QLoRA ranks first within source and last across sources. Regimes also differ in supervision amount.