What Misinformation Benchmarks Actually Measure: An Audit with VERIBENCH-WILD

Anonymous ACL submission

Abstract

When researchers report a state-of-the-art score on a misinformation benchmark, what does that number actually mean? Does the model truly recognize misleading content, or has it learned shortcuts such as post length, keywords, or patterns specific to the training data? We introduce VERIBENCH-WILD, a benchmark designed to answer these questions and audit how misinformation systems are evaluated. VERIBENCH-WILD is a multilingual corpus of 27,158 Community Notes and legacy fact-check records spanning 54 languages and a decade of content (2016–2026). The benchmark provides five evaluation tasks and one diagnostic task (T-Match), each with five-seed baselines. Our experiments reveal four main findings. First, on T-Match, a simple length-based rule outperforms five trained pair-classification models, while previously reported cross-benchmark transfer disappears when evaluated at larger sample sizes. Second, LLM judges systematically assign higher scores to outputs generated by models from the same family, inflating scores by +30.6 to +53.3 pp. Third, modern crowdsourced and legacy editorial sources differ in classifier performance by 32.9 pp, with most of the gap linked to differences in data curation rather than temporal drift. Fourth, trained classifiers and LLM judges disagree by an order of magnitude on the same text, especially on label categories where human annotators themselves show low agreement. We release the dataset, scripts, and baseline checkpoints.[1]

1 Introduction

Misinformation spreads faster and farther than corrections (Vosoughi et al., 2018), and many benchmarks have been introduced over the past decade to study it, including FEVER (Thorne et al., 2018), LIAR (Wang, 2017), FakeNewsNet (Shu et al., 2020) and UPFD (Dou et al., 2021), MuMiN (Nielsen and McConville, 2022), AVeriTeC (Schlichtkrull et al., 2023), and TSVer (Strong and Vlachos, 2025). Benchmark scores continue to improve, but higher scores do not always mean that models understand misinformation better.

Three recurring issues motivate this work. First, benchmark construction artefacts can create shortcuts that allow models to perform well using simple cues such as claim length or lexical overlap (Gururangan et al., 2018; McCoy et al., 2019; Schuster et al., 2019; Glockner et al., 2022). Second, open-ended outputs are increasingly evaluated using LLM judges, which may favour outputs from the same model family (Panickssery et al., 2024; Zheng et al., 2023). Third, pre-LLM and post-LLM datasets differ across multiple dimensions, including era, platform, programme, and topic (Koh et al., 2021; Schlichtkrull et al., 2023). As a result, a single train–test split often cannot identify the true source of a performance change.

Together, these issues raise a simple question: what does a state-of-the-art score on a misinformation benchmark actually tell us? Although these issues have been studied separately, no existing benchmark jointly examines all three under a common evaluation setting. To address this problem, we introduce VERIBENCH-WILD (§3–§4), a multilingual Community-Notes benchmark designed not only for evaluation but also for auditing what benchmark scores measure. Using this benchmark, we identify four main findings and one supporting diagnostic (§6.1–§6.6).Our contributions are:

- **A multilingual audit corpus and task suite.** VERIBENCH-WILD contains 27,158 author-disjoint Community-Notes and legacy fact-check records spanning 54 languages and 2016–2026. It includes five evaluation tasks and a T-Match diagnostic task, each with locked five-seed baselines (§3–§4).

[1] Anonymous release for review:https://anonymous.4open.science/r/VeriBenchWild-60F0.

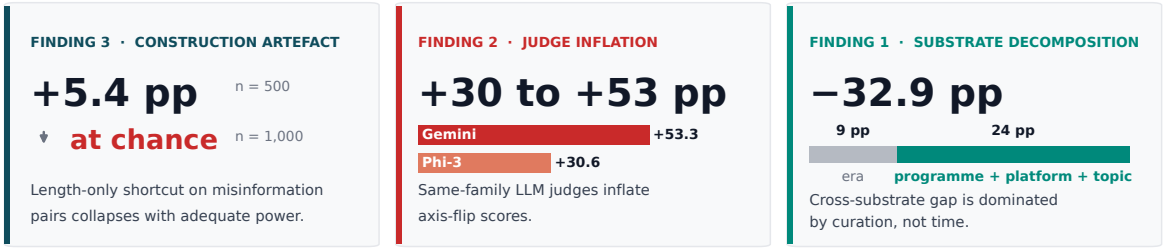

Figure 1: Three of the four headline findings of VERIBENCH-WILD (the classifier–judge disagreement finding is described in §6.5). **Left (Construction artefact, §6.4):** a length-only rule outperforms trained pair-classification models at n=500 (+5.4 pp on LIAR) but collapses to chance at n=1,000, consistent with a statistical-power artefact rather than genuine cross-benchmark transfer. **Centre (Judge inflation, §6.3):** same-family LLM judges inflate axis-flip scores by $+30.6$ pp (Phi-3-medium (Abdin et al., 2024)) to $+53.3$ pp (Gemini-2.5-Flash (Comanici et al., 2025)) relative to two independent cross-family judges scoring the same MISTRAL-7B-QLORA (Jiang et al., 2023; Dettmers et al., 2023) generations. **Right (Substrate decomposition, §6.2):** the -32.9 pp cross-substrate transfer gap decomposes into a small era component (≤ 9 pp) and a dominant programme/platform/topic residual (≥ 24 pp).

- **A statistical-power lesson for shortcut studies.** A length-based baseline outperforms trained pair-classification models in-domain, while the previously reported above-chance LIAR estimate at n=500 disappears when evaluated at n=1000 (§6.4).
- **Controlled analyses of benchmark behaviour.** We show that (i) only ≤ 9 pp of the -32.9 pp cross-substrate performance gap is attributable to era, while ≥ 24 pp is associated with programme and platform differences; (ii) same-family LLM judges systematically inflate scores; and (iii) classifiers and judges disagree most strongly on ambiguous label categories (§6.2–§6.5).
- **A reproducible two-tier release with reduced linkage risk.** The release package includes sealed splits, seed manifests, evaluation scripts, and baseline checkpoints. Tier-1 removes direct linkage keys (tweet IDs, handles, exact timestamps), while Tier-2 provides gated verbatim access (licensing in Ethics).

2 Related Work

Fact-checking and misinformation benchmarks. Automated fact-checking was introduced by Vlachos and Riedel (2014); see Guo et al. (2022) for a recent survey. Major public datasets include FEVER (Thorne et al., 2018), LIAR (Wang, 2017), LIAR-PLUS (Alhindi et al., 2018), LIAR2 (Xu and Kechadi, 2024), MultiFC (Augenstein et al., 2019), FakeNewsNet (Shu et al., 2020)/UPFD (Dou et al., 2021), MuMiN (Nielsen and McConville, 2022), MultiClaim (Pikuliak et al., 2023), and AVeriTeC (Schlichtkrull et al., 2023). TSVer (Strong and Vlachos, 2025) further provides expert-curated evaluation on 304 real-world claims using time-series evidence.

VERIBENCH-WILD complements these datasets in two ways. First, it combines 27,158 records with author-disjoint splits, propagation graphs, and AI-attribution labels, a combination not jointly available in existing datasets (FEVER/LIAR/AVeriTeC lack propagation graphs and AI-attribution; UPFD/MuMiN provide propagation graphs but no typology axes or counterfactual task; see Table 3). Second, rather than serving only as a benchmark target, it is designed as an audit framework for studying what benchmark scores actually measure.

Construction-artifact diagnostics. McCoy et al. (2019) characterized lexical-overlap shortcuts in NLI, while Gururangan et al. (2018) and Poliak et al. (2018) showed that hypothesis-only baselines can match full-input models. Niven and Kao (2019) showed that BERT solved the Argument Reasoning Comprehension Task primarily through dataset artefacts, and Schuster et al. (2019) demonstrated similar effects with claim-only FEVER baselines. These are examples of broader shortcut-learning behavior (Geirhos et al., 2020).

Finding 3 extends HANS-style shortcut analysis to misinformation pair-classification and examines whether such shortcuts transfer across benchmarks at increasing sample sizes (§6.4). The results show a strong dependence on statistical power.

LLM-as-judge protocols. Zheng et al. (2023) introduced the MT-Bench LLM-judge protocol and reported self-preference effects. Liu et al.

Glossary. Key terms used throughout the paper.

Community Notes	X/Twitter crowdsourced fact-checking programme launched in 2021.
UPFD	User Preference-aware Fake News Detection dataset (Dou et al., 2021); PolitiFact+GossipCop articles, 2016–2018.
Substrate	A source ecosystem (e.g., 'PolitiFact 2016–2018', 'Community Notes 2023–2026'). Programme, platform, era, and topic co-vary across substrates.
Six-axis typology	Six non-exclusive misleadingness categories applied per record; defined in §4 (T1 task).
Axis-flip	T0 event: a counterfactual rewrite causes a misleadingness label to change from present to absent.
Shortcut probe	Experiment testing whether a simple feature (e.g., length) solves a task; a positive result implies a construction artefact.
T-Match	Pair task: given two same-topic claims (one true, one false), decide which is true. Released as an artefact diagnostic, not a benchmark task.
T2-HP	AI-attribution task; HP = Human/AI Provenance (binary human-vs-LLM classifier).

Figure 2: Compact glossary of the eight terms used most often in Sections 3–6.6.

(2023) proposed G-Eval, Zhu et al. (2025) developed dedicated LLM judges, and Yuan et al. (2024) incorporated judge signals directly into training. Panickssery et al. (2024) later quantified self-preference across multiple tasks.

Finding 2 extends Panickssery et al. (2024) to misinformation counterfactual evaluation by testing whether self-preference remains under counterfactual axis-removal (§6.3).

Cascade dynamics. Early work studied credibility in social cascades (Castillo et al., 2011) and rumor-cascade structure (Friggeri et al., 2014; Zubiaga et al., 2016). Later, Vosoughi et al. (2018) showed that false news spreads farther than true news across several news categories.

Our cascade diagnostic (§6.6) examines whether this pattern remains consistent or varies with the curation source using the PolitiFact and GossipCop splits from UPFD.

Concept drift and distribution shift. WILDS (Koh et al., 2021) introduced a general framework for studying distribution shift, while AVeriTeC (Schlichtkrull et al., 2023) highlighted era-related shift in fact-checking datasets. Finding 1 applies a WILDS-style analysis to misinformation and further separates the transfer gap into era-related and programme-related components using two controlled experiments (§6.2).

3 The VERIBENCH-WILD Dataset

3.1 Sources and Construction

VERIBENCH-WILD combines claims from two complementary real-world sources. The first source uses Twitter/X Community Notes from 2023–2026. Notes that flag tweets as misleading become WILD records, while notes that add contextual information become WILD_SUPPORT. Notes that were later removed or marked as non-misleading form the WILD_CONTROL arm.

The WILD_SUPPORT and WILD_CONTROL tiers are used to build the T-Match diagnostic: each same-topic pair contains one WILD_SUPPORT and one WILD_CONTROL example, forming a true/false pair. The second source, UPFD_LEGACY (Dou et al., 2021), contributes PolitiFact and GossipCop cascades from 2016–2018. These records anchor the pre-LLM period and provide propagation graphs for cascade tasks.

An example record is shown in Figure 3. Each record contains five fields: (i) claim text and a salted SHA-256 author handle; (ii) evidence URLs from Community-Notes raters or PolitiFact/GossipCop annotators; (iii) snapshot age and MBFC (Media Bias/Fact Check) credibility ratings for cited domains; (iv) propagation cascades when available; and (v) human verdicts unified into three classes (SUPPORTED, REFUTED, and NOT ENOUGH INFORMATION).

The dataset spans 54 languages, with English, Spanish, Japanese, French, and Portuguese as the largest language groups. License details are provided in Appendix A (UPFD-legacy: CC-BY-SA-4.0; new tiers: CC-BY-NC-SA-4.0), and Croissant metadata (Akhtar et al., 2024) is included with the release. Table 3 compares VERIBENCH-WILD with six prior benchmarks.

3.2 Schema and Sealed Splits

Each record is represented as a JSON object with five top-level fields: `claim_text`, `evidence` (note text and cited URLs), `source` (author hash, MBFC rating, and URL domain classes), `propagation` (cascade side-car), and `labels` (six-axis typology, AI flag, three-class verdict, and 5-class severity).

We use a frozen *author-disjoint* 21,852 / 2,651 / 2,655 split, ensuring that no salted-and-hashed

id `vw-1710550235725041870` **tier** WILD **lang** en **era** post-LLM
claim *Cartoon Cow Decorative Ornament. Order here* `<URL>` *Collection* `<URL>` `<media>`
evidence note *This image is AI generated and you would therefore not receive the product as shown in the images.*
6-axis labels factualError ✓ manipulatedMedia ✓ outdatedInformation ✗ missingImportantContext ✗ unverifiedClaimAsFact ✓ satire ✗
ai_flag ✓ **verdict** REFUTED **cascade** 47-node sidecar

Figure 3: A single VERIBENCH-WILD record (test split): claim text, Community-Notes evidence, six-axis typology labels, binary AI-attribution flag, three-class verdict, and (where available) a propagation cascade. Author handles are SHA-256-hashed; URLs and media tokens abbreviated.

handle appears in more than one split. Intersections between splits are empty (verified using cohort sidecars), preventing subject-level leakage during model training. We additionally define a separate *HARD* slice ($n = 1{,}324$) from the test set to capture more challenging cases, including records without fact-check URLs, lone-token AI claims, and short non-English text. Construction details, licensing information, and task-specific subsets are provided in Appendix A.

3.3 Inter-Annotator Agreement

To assess label quality, we conducted two annotation studies. *Packet 1* (binary AI-flag) involved four PhD annotators on 190 records and yielded Fleiss $\kappa = 0.94$. Since this evaluates a single and relatively clean schema component, it is not directly comparable to multi-class verdict κ values reported for FEVER (Thorne et al., 2018), AVeriTeC (Schlichtkrull et al., 2023), or MuMiN (Nielsen and McConville, 2022).

Packet 2 (six-axis typology) used three trained annotators on a stratified sample of 50 records. Per-axis agreement is reported in Table 1. Agreement varies substantially across categories: MANIPULATEDMEDIA reaches $\kappa = 0.89$ (almost perfect), SATIRE reaches 0.60 (substantial), while FACTUALERROR, OUTDATEDINFORMATION, and UNVERIFIEDCLAIM fall between 0.38 and 0.46 (fair-to-moderate). MISSINGIMPORTANTCONTEXT drops to near chance level ($\kappa = 0.08$). Mean Fleiss κ across all axes is 0.47.

The highest-agreement categories are partly influenced by explicit keywords in Community-Notes text. This effect is strongest for AI-flag ($\kappa = 0.94$), where annotators frequently see terms such as "AI-generated" or "deepfake". A weaker version appears for MANIPULATEDMEDIA ($\kappa = 0.89$), where similar keywords are common.

By contrast, MISSINGIMPORTANTCONTEXT ($\kappa = 0.08$) and FACTUALERROR ($\kappa = 0.40$) remain difficult categories. Annotators often differed on whether any omitted but relevant fact should be labeled, or only omissions that materially change the meaning of a claim (worked examples in App. B). This highlights an important issue for evaluation: a single macro-F1 score can aggregate categories with reliability ranging from 0.08 to 0.89, while even the highest-agreement categories may partly rely on keyword cues.

Axis	H–H Fleiss κ (3 raters)	M–H Cohen κ (5-seed model)	H–Gold Mean Cohen (3 raters)
ManipulatedMedia	**0.89**	**0.82**	0.85
Satire	0.60	0.05	0.21
OutdatedInfo	0.46	0.08	0.31
FactualError	0.40	0.07	0.16
UnverifiedClaim	0.38	0.30	0.51
MissingContext	0.08	0.11	0.24
Mean	**0.47**	0.24	0.38

Table 1: Per-axis agreement on the 50-record stratified sample. Column 1 is the three-rater human Fleiss κ; column 2 is the mean Cohen κ between the 5-seed soft-voted T1 RoBERTa classifier and each human rater; column 3 is each rater's mean Cohen κ against the CN gold label. The model-human and human-human rankings correlate per axis ($\rho > 0.8$): the model agrees with humans on MANIPULATEDMEDIA (where humans agree) and fails on MISSINGIMPORTANTCONTEXT (where humans cannot agree).

Tier	Records	Era
UPFD_LEGACY	5,147	2016–2018 (pre-LLM)
WILD	5,262	2023–2026 (post-LLM)
WILD_CONTROL	8,701	2023–2026 (post-LLM)
WILD_SUPPORT	8,048	2023–2026 (post-LLM)
Total	**27,158**	

Table 2: VERIBENCH-WILD record counts by tier and era. Splits are author-disjoint and not stratified across tiers (see §3.2).

3.4 Dataset Statistics

Table 2 summarizes VERIBENCH-WILD by tier × era. Most WILD records come from the post-LLM period (2023–2026), while the pre-LLM period (2016–2018) consists entirely of UPFD_LEGACY. The transition period (2019–2022) contains relatively few records and is therefore excluded from headline analyses.

Dataset	Records	Langs	Real-world	Cascade	Author traj.	AI-attrib.	Multi-axis	Author-disj.
FEVER (Thorne et al., 2018)	185,445	1	✗	✗	✗	✗	✗	–
LIAR (Wang, 2017)	12,836	1	✓	✗	–	✗	✗	–
UPFD (Dou et al., 2021)	5,778	1	✓	✓	✗	✗	✗	–
MuMiN (Nielsen and McConville, 2022)	12,914	41	✓	P	–	✗	✗	–
AVeriTeC (Schlichtkrull et al., 2023)	4,568	1	✓	✗	✗	✗	P	–
X-FACT (Gupta and Srikumar, 2021)	31,189	25	✓	✗	✗	✗	✗	–
VERIBENCH-WILD (ours)	**27,158**	**54**	✓	P	✓	✓	✓	✓

Table 3: Dataset comparison along seven structural dimensions. ✓ = feature explicitly released; ✗ = absent-by-construction; P = partial (see text); – = not reported in the cited release.

4 Tasks

VERIBENCH-WILD includes *five evaluation tasks and one construction-artifact diagnostic* (T-Match; see Figure 2 for terminology used throughout this section). All tasks use the same sealed test split, evaluation framework, and locked five-seed baselines (Table 4; full per-baseline results are provided in App. D, Tables 19, 20, 21, 22, 23, and 24). Each task is designed to address a specific audit question, stated at the end of its description.

T1 – Six-axis typology classification. This task performs multi-label classification over the six Community-Notes MISLEADING* categories: FACTUALERROR, MANIPULATEDMEDIA, OUTDATEDINFORMATION, MISSINGIMPORTANTCONTEXT, UNVERIFIEDCLAIMASFACT, and SATIRE. The input consists of claim text concatenated with note text, and performance is measured using macro-F1. The test set contains $n_{\text{test}} = 2{,}655$ examples. *Audit question:* Can a classifier reliably distinguish the six categories, or are some inherently too ambiguous for automated labeling (§6.5)?

T2-HP – AI-attribution detection (binary, out-of-distribution). This task predicts whether a tweet was generated by an LLM using the claim text and its Community-Notes evidence. The suffix `HP` stands for Human/AI Provenance. Evaluation uses the L_HIGH (positive) and NEG (negative) subsets of the sealed test split (details in App. A). The task is intentionally out-of-distribution. Most state-of-the-art AI-text detectors are trained on synthetic datasets (e.g., RAID (Dugan et al., 2024)), whereas our labels come from Community-Notes crowd annotations and contain approximately $\sim 4\%$ residual noise. Performance is measured using AUPRC with $n_{\text{test}} = 1{,}184$. *Audit question:* Does AI-attribution generalize from clean synthetic training data to noisy real-world labels?

T0 – Counterfactual generation. This task generates a minimally edited rewrite of a Community-Notes-flagged misleading claim by removing one selected T1 axis while preserving the original topic. For example, a claim labeled as OUTDATEDINFORMATION should be rewritten so that the outdated label no longer applies. We define an *axis-flip* as a transition where a label present in the original text is absent in the rewritten version. Performance is measured using axis-flip rate judged by an independent LLM judge, with $n_{\text{test}} = 2{,}461$ generations per seed. Two open-source systems are released: `mistral7b-qlora-axis` (Jiang et al., 2023; Dettmers et al., 2023) and `qwen-0.5b-sft` (Yang et al., 2024). We report cross-family Phi-3-medium-4k-instruct (Abdin et al., 2024) (3.8B parameters) judge scores to reduce self-preference effects (§6.3). Generation and judge prompts are listed in App. E. *Audit question:* Do targeted edits change the labels assigned by classifiers and judges to the same text?

T-Prop – Cascade propagation. This task performs multi-class verdict classification (SUP / REF / NEI) using propagation graphs, with performance measured by macro-F1. We provide GAT-v2 (Brody et al., 2022), GraphSAGE (Hamilton et al., 2017), GraphGPS (Rampášek et al., 2022), and LightGBM (Ke et al., 2017) baselines. Evaluation is restricted to the UPFD_LEGACY subset ($n_{\text{test}} = 5{,}086$ cascade-bearing records out of $5{,}147$ total), since post-LLM WILD retweet trees are no longer available through X syndication. Cascade sidecars are released under gated Tier-2 access (CC-BY-NC-SA-4.0); see §7. *Audit question:* Do propagation-graph features provide predictive information beyond text in this pre-LLM fact-check setting?

T-Match – Topic-controlled helpfulness pairs (construction-artifact diagnostic). This task performs pairwise classification over same-topic claim pairs consisting of one WILD_SUPPORT and one WILD_CONTROL example. The goal is to identify the true claim. Performance is measured us-

ing pair accuracy with $n_{\text{test}} = 145$ pairs. *Released as a construction-artifact diagnostic rather than a benchmark task.* Truth labels correlate by construction, and claim length outperforms five trained architectures (Finding 3). T-Match is retained to support shortcut analysis. *Audit question:* Can simple artefacts such as claim length outperform trained models?

T-Author – Author trajectory forecasting. This task predicts whether an author's next claim is AI-generated using up to five previous Community-Notes-flagged posts and the target claim text. The history window always precedes the prediction target to prevent temporal leakage. Performance is measured using ROC-AUC; test cohort statistics are provided in App. D. *Audit question:* Can an author's previous history predict the AI provenance of their next claim?

5 Experimental Setup

Protocol: All baselines are trained on the sealed VERIBENCH-WILD v4.1.1 train split (21,852 records), selected on val (2,651), and evaluated once on the author-disjoint test split (2,655). We report five-seed ($\{7, 13, 42, 43, 44\}$) mean and standard deviation with 95% bootstrap confidence intervals (1,000 seed resamples) for headline metrics. These intervals reflect seed variability rather than test-set uncertainty. Full hyperparameters are provided in App. C.

Compute: All experiments were run on a private NVIDIA H100 80GB cluster. Total compute was approximately 120 GPU-hours for five-seed baseline training and 70 GPU-hours for diagnostic ablations; detailed breakdowns are provided in Appendix C.

Baseline coverage: Headline baselines are locked across all 5/5 seeds and frozen before analysis. Floor baselines (random, majority, length-only LR, TF-IDF) provide reference points for seed variability. Zero-shot systems (Binoculars (Hans et al., 2024), Fast-DetectGPT (Bao et al., 2024), NLI-DeBERTa (He et al., 2023; Laurer et al., 2023), and Phi-3 judge (Abdin et al., 2024)) are deterministic and reported with a single seed. We additionally evaluate Mistral-7B-Instruct (Jiang et al., 2023) as a same-family judge for Finding 2 and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as a T-Match judge (App. E). The Llama result is reported from logs and is not part of the headline Finding 2 analysis.

6 Experiments and Findings

We first present the baseline results for each task (§6.1), followed by four main findings (§6.2–§6.5) and one supporting diagnostic (§6.6). Figure 1 previews three of the four findings; the fourth (classifier–judge disagreement on ambiguous axes) is discussed in §6.5. Finding 3 provides the main methodological result by examining shortcut effects within and across benchmarks.

6.1 Baseline Results

Table 4 reports the strongest baseline for each task; complete per-seed results are provided in App. D. Full baseline results are provided in Appendix D.

6.2 Finding 1 – Cross-substrate transfer gap decomposes into era and programme axes

Training a verdict classifier (SUPPORTED vs REFUTED) on one dataset and evaluating it on another causes a large performance drop. When transferring between UPFD-legacy and Community Notes, where era, programme, platform, and topic all differ simultaneously, macro-F1 decreases by 38.0 pp for TF-IDF (95% CI $[-40.4, -35.6]$)[2] and $\mathbf{32.9 \pm 1.8}$ pp for RoBERTa. Detailed transfer results are shown in Table 5, with decomposition summarized in Table 6. We next perform two controlled experiments to determine which factors contribute to this performance drop:

(i) **LLM-boundary, programme held constant** (LIAR2 (Xu and Kechadi, 2024), PolitiFact 2000–2023). We train RoBERTa-base on records before 2022-11-30 ($n = 16{,}373$) and evaluate on records after this boundary. Performance increases by $+\mathbf{6.4 \pm 1.9}$ pp (95% bootstrap CI $[+5.1, +8.1]$; all five seeds show the same trend), indicating that post-2022 data is *easier*, not harder. This effect appears specific to changes in PolitiFact's editorial mix and does not support the hypothesis that crossing the Nov-2022 LLM boundary alone explains the observed -33 pp transfer degradation.

(ii) **Continuous intra-era drift, programme held constant** (Community Notes 2023 $\rightarrow$ 2025). We train RoBERTa-base on CN-2023 ($n = 5{,}907$) and evaluate on CN-2025 ($n = 2{,}385$). Performance changes by $+\mathbf{2.5 \pm 0.7}$ pp (95% CI $[-1.3, +6.2]$),

[2]The TF-IDF cell uses a binomial-proxy analytic (Wilson) CI; RoBERTa cells use the 5-seed paired bootstrap.

Task	Best baseline	Metric	Score (mean ± std; 95% CI)	Floor baseline
Five evaluation tasks (5/5 seeds locked)				
T1 (typology)*	RoBERTa-base (Liu et al., 2019)	macro-F1	**0.533** ± **0.002** [0.531, 0.535]	0.499 (TF-IDF)
T2-HP (AI-attrib.)†	TF-IDF + logreg	AUPRC	**0.778** (deterministic)	0.618 (length-only)
T0 (counterfactual)	Mistral-7B QLoRA (Jiang et al., 2023; Dettmers et al., 2023)	axis-flip	**0.799** ± **0.011** [0.788, 0.807]	0.493 (Phi-3 5-seed)
T-Prop (cascade)	GraphGPS (Rampášek et al., 2022)	macro-F1	**0.968** ± **0.002** [0.967, 0.970]	0.805 (TF-IDF)
T-Author (trajectory)	Transformer-history	ROC-AUC	**0.833** ± **0.001** [0.832, 0.834]	0.500 (majority)
Construction-artifact diagnostic (not a benchmark task)				
T-Match (pair)	length-longer	pair-acc	**0.579** [0.497, 0.655]	0.537 (DeBERTa-pair)

Table 4: Headline baseline per task. Scores are 5-seed mean ± std with 95% bootstrap CI; T-Match's CI is over n=145 test pairs. *On T1, the top five trained transformers are within seed-noise (range 0.523–0.533); ranking not statistically separable. †On T2-HP the trained RoBERTa attains only 0.603 AUPRC, below the TF-IDF floor; we position T2-HP as a noisy-label audit task (§7).

Model	Train	Test (macro-F1)	
		UPFD-legacy	CN (post-LLM)
TF-IDF	UPFD	**0.812**	0.336
	CN	0.396	**0.679**
RoBERTa	UPFD	**0.810** ± 0.010	0.461 ± 0.024
	CN	0.417 ± 0.016	**0.726** ± 0.006

Table 5: Cross-dataset transfer macro-F1 between UPFD-legacy and Community Notes. Bold cells indicate in-dataset evaluation (train and test from the same source), while off-diagonal cells represent cross-dataset transfer. TF-IDF results are deterministic; RoBERTa results are reported as 5-seed mean ± std. The bidirectional mean performance drop is -32.9 ± 1.8 pp. Further decomposition is provided in Table 6.

Effect (controlled experiment)	Δ F1 (pp)
Cross-substrate (UPFD ↔ CN, 5-seed)	-32.9 ± 1.8
LLM-boundary (LIAR2 PolitiFact, balanced)	$+6.4 \pm 1.9$ *(post easier)*
Continuous era (CN 2023 → 2025)	$+2.5 \pm 0.7$ *(CI incl. zero)*
Era axis combined (upper bound)	≤ 9
Programme/platform/topic residual (lower bound)	≥ 24

Table 6: Finding 1 decomposes the -32.9 pp UPFD↔CN transfer gap. Era upper bound sums absolute values ($|+6.4| + |+2.5| \approx 9$ pp). Both era controls move opposite to the -32.9 pp gap (post-2022 is *easier*, not harder), so the true era contribution to degradation is near zero; the ≥ 24 pp programme residual is conservative.

with confidence intervals overlapping zero, suggesting no measurable effect from two years of continuous post-LLM drift.

Together, these experiments suggest that era contributes at most $\sim$9 pp to the observed transfer gap. Since one control moves in the opposite direction from the expected "LLMs degrade content" hypothesis, the remaining $\geq$24 pp is treated as a residual effect associated with programme, platform, topic, and annotation differences between UPFD-legacy and Community Notes (Table 6). These factors co-vary in the current data and cannot be further separated. Per-seed and bootstrap details are provided in App. D.

6.3 Finding 2 – LLM-judge family bias

Following Panickssery et al. (2024); Zheng et al. (2023); Liu et al. (2023), we re-evaluated T0 counterfactuals generated by Mistral-7B-QLoRA and Qwen-0.5B using the cross-family *Phi-3-medium-4k-instruct* judge. Same-family judges substantially inflate absolute axis-flip scores: inflation reaches $+30.6$ pp for Mistral and $+29.2$ pp for Qwen. However, the relative ranking between models remains stable: the Mistral–Qwen gap is $+8.9$ pp under cross-family Phi-3 and $+10.9$ pp under same-family Mistral. This suggests that LLM judges mainly distort absolute scores rather than relative ordering (Table 7).

Frontier-judge replication. We further re-evaluated the same five-seed Mistral generations using *Gemini-2.5-Flash* (cross-family frontier; n=12,305, $\kappa_{\text{PHI3,GEMINI}}$=0.39), yielding macro axis-flip 0.266 ± 0.005. The same-family effect remains, but its magnitude depends on the judge: Phi-3 estimates $+30.6$ pp inflation, whereas Gemini estimates $+53.3$ pp. The resulting 22.7 pp difference is concentrated on conceptually ambiguous axes (App. D).

6.4 Finding 3 – Length shortcut effects within and across benchmarks

In-domain. A simple length-based heuristic (predict the longer claim as truthful) achieves pair accuracy of 0.579 on T-Match (n = 145), outperforming all five trained pair-classification models (best DeBERTa-v3-large: 0.537 ± 0.028; RoBERTa-base: 0.524 ± 0.032; full results in App. D). However, performance drops to chance on within-source UPFD pairs (0.46 on GossipCop and 0.28 on PolitiFact), indicating that the shortcut reflects a HANS-style (McCoy et al., 2019) pair-construction artifact rather than a true veracity signal. *Cross-benchmark.*

Judge	Mistral-7B	Qwen-0.5B
Absolute axis-flip rate (fraction of rewrites credited)		
Mistral-7B (Jiang et al., 2023) (same family)	0.799 ± 0.011	0.690 ± 0.019
Phi-3-medium (Abdin et al., 2024) (cross)	0.493 ± 0.009	0.398[a]
Gemini-2.5-Flash (Comanici et al., 2025) (cross)	0.266 ± 0.005	0.198 ± 0.013
Same-family inflation (same-family rate − cross-family rate)		
vs. Phi-3 reference	+30.6 pp	+29.2 pp
vs. Gemini reference	+53.3 pp	+49.2 pp

Table 7: Cross-family T0 axis-flip judge comparison. Upper block: macro axis-flip rate (fraction of rewrites the judge scored as removing the targeted axis). Lower block: same-family inflation. Mistral columns are 5-seed; Qwen columns 4-seed (seed 42 excluded). Gemini-2.5-Flash: n=12,305 judgments, $\kappa_{\text{PHI3,GEMINI}}$=0.39. Gemini–Mistral inflation 95% CI $[+52.2, +54.4]$; per-seed values in App. D.

[a]Single Phi-3 seed on the Qwen arm.

Benchmark	n=500	n=1,000 [CI]	Verdict
VERIBENCH-WILD T-Match*	**0.579**	**0.579** [0.497, 0.655]	above chance
LIAR (random)	0.554	0.521 [0.489, 0.551]	at chance
AVeriTeC (random)	0.484	0.509 [0.479, 0.540]	at chance

Table 8: Length-longer-is-true heuristic at increasing sample size. Metric: pair-accuracy (chance = 0.500). CIs are paired bootstrap (1,000 resamples) at n=1,000; the original n=500 probe reported point estimates only. The LIAR verdict flips from above chance (+5.4 pp) to at chance purely from doubling n, consistent with a statistical-power artefact rather than genuine cross-benchmark transfer. On within-source UPFD pairs the heuristic degrades to chance (GossipCop 0.46, PolitiFact 0.28 at n=500). *T-Match fixed at n=145 pairs.

At $n = 1{,}000$, confidence intervals overlap with chance for both cross-benchmark evaluations (Table 8). The previously observed LIAR result at $n = 500$ (+5.4 pp above chance) disappears at the larger sample size, suggesting a statistical-power effect rather than genuine transfer (§7).

6.5 Finding 4 – T1 classifier is not axis-causal

For each T0 generation targeting axis a, we apply both the T1 classifier and the Phi-3 judge to the original and rewritten text. The T1 targeted-axis flip rate (TFR) is 0.031 ± 0.003 for Mistral-7B and 0.039 ± 0.003 for Qwen-0.5B, compared with Phi-3 TFR values of 0.49 and 0.40, corresponding to a 16×/10× classifier–judge disagreement. To test whether this difference reflects classifier under-training, we evaluate an independent zero-shot NLI baseline (DeBERTa-v3-large-mnli-fever-anli) on the same records. Its macro TFR (0.083) is much closer to the trained T1 classifier (∼0.03) than to Phi-3 (0.49). Together with the human agreement ordering in Table 1, three independent methods agree: humans, the trained classifier, and the NLI baseline all suggest that the targeted axis did not flip, while only the LLM judge predicts otherwise (App. D.7).

6.6 Supporting Diagnostic

We include one additional probe on cascade behavior using UPFD. Two supporting analyses are deferred to App. D: per-language T1/T2-HP results (English is not the easiest language despite 81% training share) and a frozen-probe cross-task transfer matrix (the T1 encoder does not positively transfer downstream).

Curation-conditional cascade-shape signature on UPFD. On the PolitiFact/GossipCop split of UPFD, the cascade pattern reported by Vosoughi et al. (2018) depends on the curation source. POLITIFACT ($n = 284$) reproduces the original finding (REFUTED spreads farther; $\delta = -0.15$), whereas GOSSIPCOP ($n = 4{,}871$) shows the opposite trend ($\delta = +0.61$). Since UPFD is approximately 17:1 GossipCop-dominated, aggregate analysis reports the inverted direction; full statistics are provided in App. D.

7 Discussion and Conclusion

Our findings suggest two broad challenges in misinformation evaluation. The first is *benchmark construction*. Finding 3 shows that simple shortcuts such as claim length can outperform trained models within a dataset yet fail to transfer across datasets at larger sample sizes. The second is *evaluation itself*. Findings 2 and 4 show that same-family LLM judges can inflate absolute scores and disagree strongly with both classifiers and human judgments on ambiguous categories. Finding 1 places these effects into a broader setting using a WILDS-style decomposition (Koh et al., 2021). Of the observed −32.9 pp transfer gap, at most ≤ 9 pp is associated with era, while the remaining ≥ 24 pp is linked to differences in programme, platform, and topic.

These observations motivate three practical recommendations for misinformation evaluation: (i) report simple baselines such as length-only and TF-IDF for pair-construction tasks; (ii) evaluate shortcut transfer with sufficient statistical power ($n \geq 1{,}000$); and (iii) report absolute LLM-judge inflation using at least two cross-family judges. VERIBENCH-WILD (v4.1.1; CC-BY-NC-SA-4.0) is released as a reusable benchmark and diagnostic framework for more reliable misinformation evaluation.

Limitations

Dataset coverage. VERIBENCH-WILD inherits the Community-Notes content distribution and therefore over-represents posts receiving sufficient flag activity for review. English dominates the corpus (81 %, n=2,095 test); Spanish is the next-largest language (n=154), while all remaining languages have n<75 test examples and are underpowered for reliable per-language conclusions. T1 macro-F1 aggregates six axes with substantial variation in label reliability, with per-axis Fleiss κ ranging from 0.08 (MISSINGCONTEXT) to 0.89 (MANIPULATEDMEDIA) (Table 1). The WILD stream spans 2023–2026 and UPFD_LEGACY spans 2016–2018; the 2019–2022 transition period is sparse and excluded from headline analyses. A survivor-selection bias also exists: only 35.6 % of attempted hydrations succeeded (App. A), biasing the sample toward content still accessible in 2026-05. Results are not explicitly rebalanced by topic, geography, or protected groups.

Post-hoc cascade reanalysis. The curation-conditional cascade signature (§6.6) is a post-hoc analysis of the published UPFD PolitiFact/GossipCop split rather than a pre-registered study on independent cascades. Replication on independently curated datasets is needed before treating the observed direction reversal as a general property of editorial and entertainment fact-check sources.

Era-decomposition co-confounds. The pre/post-Nov-2022 PolitiFact comparison keeps the programme fixed but not the claim-topic distribution. The post-2022 editorial mix shifted toward more viral social-media claims with stronger lexical cues that may benefit the model. The observed +6.4 pp result therefore reflects changes in claim mix rather than evidence that post-LLM misinformation is intrinsically easier. The intra-Community-Notes comparison also covers only approximately two years of drift.

T2-HP scope and AI-flag. We position T2-HP as a noisy-label robustness task rather than a clean AI-detection benchmark. The AI_FLAG labels rely on Community-Notes annotators identifying AI-generated content, and large-scale recall and precision are not directly measured. High agreement on the binary AI-flag (§3.3) and MANIPULATED-MEDIA partly reflects explicit cues such as "AI-generated" and "deepfake" in positive-class notes. To estimate dependence on these cues, we retrained RoBERTa with trigger keywords masked: AUPRC decreased by only 3.2 pp (95% CI $[-4.9, -1.6]$; $0.603 \rightarrow 0.571$). This suggests that the model relies on broader lexical patterns beyond explicit keywords. However, the masked result remains below the length-only baseline (0.618, Table 4), suggesting limited semantic understanding. Finding 4 (§6.5) suggests that similar axis-causality limitations may also apply to the AI-flag.

LLM-judge protocol. T0 absolute axis-flip scores depend on the judge model (§6.3). We use cross-family Phi-3 as the reference judge, while same-family Mistral results are reported only for comparison. Finding 2 (§6.3; Table 7) shows that score inflation depends on the reference judge: Phi-3-medium-4k-instruct estimates +30.6 pp inflation, while Gemini-2.5-Flash estimates +53.3 pp. The directional effect is consistent, although its magnitude may vary across judges. The Qwen-0.5B arm reproduces the same trend under Gemini (+49.2 pp, 4 seeds).

Coverage and propagation gaps. Propagation graphs are available for all UPFD_LEGACY records (n=5,086) and a subset of WILD (n=5,352); the remaining records are node-only. Cascade-based analyses (§6.6, App. D) are therefore limited to graph-available subsets. A preprocessing artifact affected 21 TRAIN/VAL records (**none in test**), where the contracted form `user's` collapsed into `user'<hash>` during SHA-256 handle hashing. This issue does not affect headline results; details are provided in App. A.2.

T-Match cross-tier construction confound. T-Match pairs are constructed from WILD_SUPPORT (true) and WILD_CONTROL (false) records, creating a perfect correlation between tier and label. Claim length therefore becomes a correlated signal, allowing a length-only heuristic to achieve 0.579 pair accuracy (Finding 3; §6.4). We report this as a diagnostic finding rather than hiding it through re-splitting. A length-stratified T-Match v2 is planned for a future release.

Future work pointer. The programme/platform/topic residual ($\geq$ 24 pp; Table 6) remains unresolved and motivates future controlled studies. One natural next step is applying Community-Notes-style annotation to PolitiFact-era content (or vice versa) to separate programme effects from topic effects.

Ethics Statement

Privacy and linkage risk. VERIBENCH-WILD is derived from publicly available Community-Notes content and salts SHA-256 author handles. A linkage probe on the sealed test split found that 81.0 % of records (Wilson 95% CI $[78.2, 83.5]$) remain re-identifiable through combinations of public metadata. Since hashed handles alone are not a strong anonymity mechanism, we disclose this limitation explicitly and mitigate it through a two-tier release strategy described below.

Author consent and IRB status. The WILD stream consists of publicly available Community-Notes-flagged posts collected under X platform terms. Individual consent was not feasible at dataset scale. Under our institution's policy on secondary use of public social-media content, the dataset and annotation process did not require IRB review. Additional rationale is provided in App. G. A content-removal and takedown process will be documented in the published data card (App. A).

Annotators and minors. The 6-axis annotation packets (§3.3) were produced by PhD-level researchers from the authors' academic network in a research-collaboration setting rather than through paid crowd work. Annotator demographics and dispute-resolution procedures are described in App. B. We did not apply a systematic minor-content filter; a manual review of annotation packets (240 records total) did not identify content involving minors.

Dual use. VERIBENCH-WILD could potentially support the study of methods for generating platform-evasive misinformation. However, the released T0 generation systems (Mistral-7B-QLoRA-SFT and Qwen-0.5B-SFT) are small open-source models with widely available alternatives, and we consider the additional risk small relative to the benefits of transparency and reproducibility.

Carbon cost. All experiments together required approximately 120 GPU-hours on H100 80GB nodes, corresponding to an estimated energy use of ~ 72 kWh or ~ 29 kg CO_2-eq.

Linkage mitigation and release policy. We mitigate privacy risks through a two-tier release. Tier-1 removes identifiers such as `tweet_id`, raw handles, timestamps, and direct URLs while retaining task-relevant annotations. Tier-2 includes verbatim records and cascade sidecars required for T-Prop under gated CC-BY-NC-SA-4.0 access. Researchers requiring stronger anonymity should use Tier-1 or synthetic Community-Notes data. Access procedures and takedown details are provided in App. A.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Luca Foschini, Pieter Gijsbers, Joan Giner-Miguelez, Sujata Goswami, Nitisha Jain, Michalis Karamousadakis, Satyapriya Krishna, and 1 others. 2024. Croissant: A metadata format for ml-ready datasets. volume 37, pages 82133–82148.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *Preprint*, arXiv:2311.16867.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

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

Shaked Brody, Uri Alon, and Eran Yahav. 2022. How attentive are graph attention networks? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pages 675–684.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

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

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

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

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. 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)*, pages 12463–12492.

Adrien Friggeri, Lada A. Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 101–110.

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

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

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling (COLM)*. ArXiv:2312.00752.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-FACT: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP), Short Papers*, pages 675–682.

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

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Alan G. Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. ArXiv:2111.09543.

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. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pang Wei Koh and 1 others. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

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

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *Preprint*, arXiv:2312.17543.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. 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)*, pages 2511–2522.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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)*, pages 3428–3448.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 24950–24962.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 220–229.

Dan Saattrup Nielsen and Ryan McConville. 2022. 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)*, pages 3141–3153.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4658–4664.

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

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16477–16500.

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. 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*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. 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.

Marek Strong and Andreas Vlachos. 2025. Tsver: A benchmark for fact verification against time-series evidence. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29894–29914.

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

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

William Yang Wang. 2017. "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*, pages 422–426.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Cheng Xu and M-Tahar Kechadi. 2024. An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12:88006–88021. LIAR2 dataset; also distributed via Hugging Face: `chengxuphd/liar2`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, and 1 others. 2024. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pages 51257–51296.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3):e0150989.

A Dataset Details

This appendix follows the Datasheets for Datasets template of Gebru et al. (2021), organised as prose under the standard Gebru categories. We additionally adopt the complementary documentation standards of Model Cards (Mitchell et al., 2019) and Data Statements (Bender and Friedman, 2018) where they apply to the released baselines and language-resource description. Numerical facts are anchored in on-disk reports shipped with the release (`RELEASE_VERIBENCH_WILD/data/v4_1_clean_full/`); where a prompt does not apply we say so rather than skip it.

A.1 Motivation

VERIBENCH-WILD-Wild was created to audit whether current text-only and graph-augmented misinformation classifiers actually use the signals their architectures advertise (content, evidence, source credibility, propagation, author history, generator attribution). Public benchmarks either expose curated fact-check claims in isolation (LIAR (Wang, 2017), AVeriTeC (Schlichtkrull et al., 2023), PUBHEALTH (Kotonya and Toni, 2020)) or post-hoc engagement graphs without AI attribution (FakeNewsNet (Shu et al., 2020), UPFD (Dou et al., 2021)), and none simultaneously pair real-world *AI-generated* misinformation with the evidence used to refute it, the credibility of its source, and the propagation footprint on the platform that hosted it. The dataset fills that gap and explicitly supports the four headline findings and one supporting diagnostic reported in §6.

The dataset was assembled by the authors of this paper under an academic-research role at *[anonymous institution]*, funded only by internal research compute. No commercial sponsor or funding body imposed editorial constraints. The release is a re-organisation and enrichment of three already-public sources – the X Community Notes weekly dump, the X public syndication CDN, and UPFD (Dou et al., 2021) – and introduces no new web crawl.

A.2 Composition

Instances and scale. Each instance is either one Community-Notes-flagged tweet from X (tiers `wild`, `wild_control`, `wild_support`) or one FakeNewsNet-via-UPFD record with full retweet cascade (`upfd_legacy`). Records are JSON objects with a uniform top-level schema across tiers (`id`, `source_dataset`, `tier`, `split`, `tweet_id`, `claim_text`, `claim_text_clean`, `evidence`, `source`, `propagation`, `labels`, `metadata`, `temporal`); the six MISLEADING* axes inside `labels.sub_reasons` are listed with their agreement and headline T1 F1 in Table 10 (App. B.5). The sealed v4.1.1 release contains **27,158** records[3] partitioned into **21,852** train, **2,651** validation, and **2,655** test (root files `veribench_wild_{train,val,test}.jsonl`), with per-tier counts of `wild` 5,262, `wild_control` 8,701, `wild_support` 8,048, and `upfd_legacy` 5,147. Eleven per-task subset directories ship with their own `{train,val,test}.jsonl` files derived from the root release.

Sample, not population. The wild stream is the intersection of the weekly Community Notes public dump (snapshot 2026-04, covering 2023-01 to 2026-04 in the wild tier and 2021-01 to 2026-04 across all tiers) with a hardened v2 balanced AI-attribution regex (95.7% precision on a 100-note held-out re-sample). **35.6%** of AI-positive tweets were deleted or withheld at hydration time and are absent from the release; this survivor bias is documented in the release Ethics Statement and disclosed under Limitations. The `upfd_legacy` tier is the entire FakeNewsNet/UPFD release (5,147 cascade-bearing records after URL-noise filtering, vintage 2010–2018).

Per-record content. Each instance contains preprocessed but otherwise raw natural-language text plus structured metadata: the claim text (`claim_text`), a lower-cased deduplication key (`claim_text_clean`), the Community Note text and the cited-URL bag with per-domain class tags (train-only after Fix N3, the cited-URL verdict-leak strip defined in the build pipeline below), an MBFC-derived source-credibility score, engagement counts (`favorite_count` and `conversation_count` for wild tiers; full cascade JSON for `upfd_legacy`), and a labels block (binary verdict, 3-class verdict, 5-class severity, 6-axis typology, `ai_flag` with mode and confidence). No images, video, or audio.

Labels and agreement. Multiple targets support the task lineup: the 6-axis typology

[3] Pre-dedup build count was 27,336; the 178-record delta from cross-split/intra-train duplication removal is documented in App. A.2 alongside the 21 records carrying a `user'<hash>` tokenization artifact. Both are properties of the locked v4.1.1 release.

(labels.sub_reasons), labels.ai_flag for T2-HP attribute, labels.unified (binary) and verdict_3class, the 5-class severity_5class, cascade-derived verdicts for T-Prop, future-claim labels for T-Author trajectory, and the narrative cluster pair-id for T-Match. Inter-annotator agreement is reported in full in §3.3; headline numbers are repeated here for datasheet self-containment. *Packet 1* (binary AI-flag axis): **Fleiss** $\kappa = 0.94$ on $n = 190$ records, four PhD raters. *Packet 2* (six-axis typology): three trained raters on $n = 50$ stratified records, three-rater Fleiss κ per axis ranging from 0.08 (MISSINGIMPORTANTCONTEXT) to 0.89 (MANIPULATEDMEDIA), mean 0.47 across the six axes (full per-axis κ table in App. B.5).

Schema reference. Table 9 lists the canonical per-record field structure. Every record in veribench_wild_{train,val,test}.jsonl conforms to this schema; per-task subsets under author_trajectory/, topic_matched/, etc. inherit the top-level schema and add task-specific prediction targets (documented in each subset's metadata file).

Missing fields. Several fields are absent by design. X syndication no longer exposes retweet_count, quote_count, view_count, follower counts, profile bios, or account ages, so these are declared null on wild-tier records. The cascade graph is populated only on upfd_legacy; wild-tier propagation is restricted to favorite and conversation counts. After the cited-URL verdict-leak strip (Fix N3) the evidence-cited-URL and source-linked-domain fields are removed from the val/test splits of root and upfd_legacy_only/ while train retains them. After the tier-leak strip (Fix N1) the eval splits of media_ai_attribution{,_paraphrased}/ have tier, tier_vintage, and source_dataset stripped from the top level and mirrored under metadata.* for provenance only.

Relationships and splits. metadata.narrative_cluster_id groups same-topic true/false claim pairs for T-Match; source.handle (hashed) groups per-author trajectories for T-Author; metadata.tweet_id_duplicate_count surfaces tweets with more than one CN note. Cascade edges for upfd_legacy live in a per-record .npz side-car referenced by propagation.cascade.graph_file. The primary split is an **author-disjoint** group-shuffle (group key: tweet_id ∪ source.handle; seed = 42; 80/10/10) shipped as veribench_wild_{train,val,test}.jsonl. A separate **temporal holdout** (temporal_holdout/; train $\leq$ 2024-06-30, test $\geq$ 2024-10-01) ships for distribution-shift evaluation. A sealed-leaderboard reservation withholds 20% of every test split; only the 80% portion is distributed in this release. The released sealed test (veribench_wild_test.jsonl, 2,655 records) has SHA-256 6e6e42e7497bc61cc4ed1dc517d24e9120 90a86c888066dee70056493500c989; per-file hashes for all subset splits are listed in the SHA-256 manifest shipped with the release.

Known errors and noise. Three operational issues are tracked in the release errata. (i) The AI-flag regex carries ∼4% residual noise (95.7% precision). (ii) The pre-dedup build count (27,336) differs from the locked total (27,158) by 178 records removed in the N4 cross-split and intra-train dedup pass; the paper, Croissant manifest, and SHA-256 manifest all reference the post-dedup figure. (iii) 21 train+val records (16 train, 5 val, **0 test**) contain the contracted token user's collapsed into user'<hash> from a handle-hashing regex overfire; test splits are unaffected, and the v4.1.2 errata will ship a clean-text filter that surfaces and removes them. The two paper-level concerns commonly raised against Community-Notes-based corpora — linkage re-identifiability (§7, §7) and the T-Match topic-collapse construction (§6) — are disclosed in the body and not repeated here.

External resources. The Tier 2 (gated) release is self-contained. The Tier 1 (public) release ships tweet_ids plus a hydration script (under scripts/ in the public release) that calls the public X syndication CDN endpoint cdn.syndication.twimg.com/tweet-result. The upfd_legacy cascade side-cars are bundled with the release. No paywalled or authenticated APIs are required.

Confidentiality and sensitivity. The release contains no DMs, no protected-account tweets, and no withheld-jurisdiction content; the syndication CDN refuses to return non-public content. Author handles are SHA-256 hashed with a per-release salt (kept private; hashing audit ships with the release) and never re-distributed in the clear. The wild tiers do contain content that may be offensive or anxiety-

Field	Type	Description
Top-level identifiers		
id	str	Stable record id, prefixed by tier (vw-, vw-control-, etc.)
source_dataset	str	veribench_wild or upfd_legacy
tier	str	wild / wild_support / wild_control / upfd_legacy
split	str	train / val / test
tweet_id	str	Original X tweet id (canonical join key)
claim_text	str	Verbatim tweet body
claim_text_clean	str	Lower-cased dedup key
evidence *(CN note text + cited URLs)*		
.note_text	str	Verbatim CN rater note
.note_classification	str	CN platform verdict label
.trustworthy_sources_flag	bool	CN rater's source-quality vote
.n_cited_urls	int	Count of evidence URLs
.evidence_quality_score	float	Aggregated MBFC-derived score
source *(author + credibility)*		
.handle	str	Salted SHA-256 of X handle
.is_blue_verified	bool	X paid-verified flag
.credibility_score	float	MBFC-derived author score
.prior_cn_count	int	Number of prior CN flags on this author
propagation *(cascade summary; sidecar on* upfd_legacy*)*		
.favorite_count	int	Likes at snapshot time
.conversation_count	int	Replies + quotes at snapshot
.observed_at_iso	str	UTC timestamp of snapshot
labels *(six-axis typology + verdict + AI flag)*		
.unified	str	Binary verdict (SUPPORTED / REFUTED)
.verdict_3class	str	Three-class verdict (adds NEI)
.ai_flag	bool	T2-HP target: AI-generated content
.ai_flag_mode	str	balanced / negative_control / recall
.ai_flag_confidence	float	Regex-derived precision estimate
.sub_reasons.misleading*	bool	6-axis typology (T1 target) — factualError, manipulatedMedia, outdatedInformation, missingImportantContext, unverifiedClaimAsFact, satire
metadata *(provenance + slicing keys)*		
.lang	str	ISO-639 language code (54 distinct)
.era_partition	str	pre_llm_2016_2018 / transition_2019_2022 / post_llm_2023_2026
.narrative_cluster_id	int	T-Match topic-cluster grouping key
.tier_vintage	str	Internal source-tier provenance
.handles_hashed	bool	Author-handle anonymisation status
temporal *(time + latency)*		
.tweet_created_at_ms	int	Unix-ms timestamp of original tweet
.note_created_at_ms	int	Unix-ms timestamp of CN note
.detection_latency_hours	float	Tweet→note delay
.snapshot_age_days	float	Days since tweet at observation

Table 9: Canonical per-record schema. Top-level groups are id, claim_text, evidence, source, propagation, labels, metadata, temporal. Sub-fields are indented with the leading . prefix. null fields are listed but omitted from this table; the full JSON schema ships as a Croissant manifest in the release.

inducing – political claims, AI-fabricated imagery descriptions, public-figure caricatures, pandemic-era health misinformation – which we do not redact, since detecting it is the benchmark's purpose. Sensitive topics (political opinion, religious belief, public-health claims, allegations against named public figures) also appear; researchers inherit downstream responsibility for any application-level fairness audit.

Subpopulations and identifiability. The dataset does not assign membership labels, but implicit subpopulations are inferable from the public CN data (political affiliation is often public; language is provided in metadata.lang; geographic origin can sometimes be inferred from CN context). **Individuals are identifiable by linkage attack.** 81% of test records and a similar share of train/val carry the (tweet_id, temporal.tweet_created_at_ms, source.handle) triple, and tweet_id alone is sufficient to re-hydrate the original public tweet (and therefore the original handle) via the X API. We do not redistribute raw handles, but we cannot prevent re-identification because the underlying CN dump is itself indexed by public tweet_id. Hashing prevents *offline* cross-record linkage on this release alone; it does not prevent *online* linkage against X. This trade-off is the irreducible cost of using a Community-Notes-based corpus rather than a synthetic one and is disclosed in the release Ethics Statement and the Gated Access Agreement.

Tier-1 de-identified release (shipped with this submission). The public Tier-1 split is included in the submission release at `v4_1_tier1_distilled/`. It retains the research-essential fields (claim text, community-note evidence text, 6-axis typology labels, verdict, cascade summary statistics, language, and a deterministic anonymous author identifier that enables T-Author trajectory forecasting). It strips the direct linkage vectors documented in the Ethics statement (§7): `tweet_id`, raw and hashed handles, exact timestamps (replaced with month buckets), and `t.co`/twitter.com/x.com URLs inside text (replaced with the token `[URL]`). Record counts match the locked splits (21,852 train, 2,651 val, 2,655 test). This split supports T0, T1, T-Match, T-Author, T2-HP, and all paper-reported findings; T-Prop researchers requiring full cascade graphs will be able to apply for Tier-2 gated access at `v4_1_clean_full/` starting at EMNLP 2026 camera-ready, once the institutional support contact and gated-access agreement are made public. This shifts the linkage-attack response from disclosure-only to disclosure + shipped mitigation.

A.3 Collection Process

Sources. All data is **observed and derived only – no validated direct human annotation by the authors**. Five upstream channels feed the build: (i) the X Community Notes public dump (CC0-equivalent per the X open-data download guide on `communitynotes.x.com`, under-the-hood section); (ii) the public X syndication CDN for tweet text and favorite / conversation counts; (iii) the Idiap MBFC factuality mirror for per-domain credibility buckets; (iv) UPFD (Dou et al., 2021) (CC-BY-SA-4.0) for the pre-LLM cascade tier; and (v) a Mistral-7B (Jiang et al., 2023) LLM judge for axis-flip verification of the counterfactual triples (T0). All five are public; no scraping, no authenticated API, no rate-limit bypass.

Procedures. The pipeline is Python scripts under `code/wild/` (v2 balanced regex, syndication hydrator, MBFC join, evidence enrichment, narrative clustering via embedding + UMAP (McInnes et al., 2018) + HDBSCAN (Campello et al., 2013), axis-flip generation) and `scripts/` (split builders, hashing), validated by a truncation regression suite and four formal audit reports shipped with the release.

Sampling. Wild-tier candidates are every CN-flagged tweet whose note text matches the v2 balanced AI-attribution regex (deterministic). `wild_control` is a random subsample of non-AI CN notes, hardness-matched on month bucket (Bhattacharyya overlap 0.94) and summary length relative to the wild tier. `wild_support` is any-classification CN notes added for typology breadth. `upfd_legacy` is the entire UPFD release minus URL-noise records.

Personnel and compensation. Annotator recruitment, qualifications, and compensation protocol are documented in App. B.1; upstream Community-Notes contributors are governed by the X platform.

Timeframe. The Community Notes snapshot is dated 2026-04 and covers notes filed between 2021-01 and 2026-04. Hydration ran from 2026-04 to 2026-05. The UPFD tier captures cascades from 2010 to 2018 (median 2017). The frozen v4.1.1 release was locked on 2026-05-16.

Ethical review and consent. No IRB was solicited because the data is exclusively public posts and public Community Notes, processed under the platforms' terms of use. We adopted the most conservative protocol we could: hashed handles, a removed-id takedown mechanism, a gated tier for verbatim text, and a public tier with hydration-on-demand (full protocol in the release Ethics Statement). Collection was third-party only (X, MBFC mirror, UPFD authors); tweet authors were not separately notified, because the data was published by X's own Community Notes system, to which posters consented under the X Terms of Service. A takedown channel (institutional contact email to be added on publication) will accept removal requests within seven days, and revocation will run through the same address with a `tweet_id`: the next quarterly snapshot excludes the id, and its presence is recorded in a removed-ids companion file shipped with each future release.

A.4 Preprocessing, Cleaning, Labeling

Pipeline. The build applies the following nine stages in order:

1. **Regex hardening** — v2 balanced AI-attribution regex, 95.7% precision.
2. **Multi-note collapse** — when multiple CN notes target the same tweet, prefer MISINFORMED > trustworthy-sources-flagged > earliest.

3. **Schema unification** (Fix N2) — 12-axis `labels.sub_reasons` dict[bool|None] across all tiers so PyArrow infers a single column; UPFD list-form tokens mapped to the typology.
4. **Tier-leak strip** (Fix N1).
5. **Cited-URL verdict-leak strip** (Fix N3).
6. **Cross-split and intra-train dedup** (Fix N4) — normalised `claim_text_clean` via lowercase + whitespace-collapse + MD5; dropped 44 cross-split, 127 intra-train, and 7 additional intra-train duplicates; ID lists ship with the dedup audit trail.
7. **Handle hashing** — SHA-256 with per-release salt.
8. **Counterfactual generation** for T0 — LLM-rewrite + Mistral-7B axis-flip judge.
9. **Narrative clustering** for T-Match — embedding → UMAP → HDBSCAN.

Raw data and software. The project archive preserves raw data alongside the release: `v4_clean`, `v4_clean_full`, and `v4_1_clean_full` (the released version) sit on disk with their build manifests, and the v4 foundation preserves the upstream CN snapshot and v4-era enrichment. The public release ships only the v4.1.1 snapshot to keep download size manageable; prior snapshots are available on request. All preprocessing scripts, the unified evaluation harness, and the HuggingFace builder script ship in the public release, alongside a reproduction recipe and the regenerated Croissant manifest.

A.5 Uses

Existing and supported tasks. The dataset underlies the four headline findings and one supporting diagnostic in §6, covering the five evaluation tasks (T1 typology, T2-HP attribute, T0 counterfactual generation, T-Prop cascade verdict, T-Author trajectory forecasting) plus the T-Match topic-controlled construction-artefact diagnostic. Findings also cover temporal generalisation, cross-task transfer, and cross-benchmark probes. The HuggingFace dataset card will maintain a "Papers using this dataset" section once external use appears; we commit to curating it for at least two years post-publication. The release also supports source-credibility modelling without follower counts, cross-lingual transfer studies (54 distinct `metadata.lang` values, of which 12 carry ≥ 50 records and 3,585 records carry a langdetect-verified native-language confirmation), time-to-detection regression, engagement / virality regression, LLM-judge calibration studies (T0 ships a frozen Mistral-7B judge), and narrative-cluster drift studies.

Composition-driven caveats. Six user-facing impacts are flagged as falsifiable warnings. (a) **English-skewed:** 81% of records are English; the long language tail is thin (es 4.0%, ja 2.4%, pt 2.3%, fr 2.1%, de 1.3%, the rest <1% each). Multilingual claims should be evaluated on the 3,585-record native-verified subset, not the full release. (b) **Survivor bias:** 35.6% of AI-flagged tweets were deleted/withheld at hydration, so detection-rate numbers under-estimate true platform prevalence. (c) **Cascades only on the pre-LLM tier:** T-Prop's 0.88 F1 is on the 5,147-record `upfd_legacy` subset and does *not* generalise to wild tiers, since X syndication no longer exposes retweet trees. (d) **T-Match construction confound:** `narrative_cluster_id` is both a control variable and a feature, so feeding cluster-id as input trivialises the task. We report T-Match purely as a topic-collapse and length-shortcut diagnostic (Finding 3); absolute T-Match pair-accuracies should not be read as benchmark-grade veracity scores. (e) **AI-flag is regex-distant:** `ai_flag` derives from a 95.7%-precision regex over CN note text, not from per-tweet expert annotation, so T2-HP is a noisy-label audit task rather than a clean AI-detection benchmark. (f) **Cited-URL verdict leakage is removed from val/test but retained in train,** so models that condition on cited URLs at training time will appear to generalise; users should re-train without those fields for any deployment claim.

Out-of-scope uses. The dataset **should not** be used to: (i) train generators conditioned on `ai_flag = True` examples to produce content stylistically similar to flagged misinformation; (ii) identify or de-anonymise individual CN raters (their identifiers are pseudonymous in the upstream dump and we do not attempt to invert that); (iii) make automated moderation decisions against accounts without human-in-the-loop review; or (iv) inform clinical, legal, hiring, lending, or any other consequential individual-level decision-making. The dataset is designed as a diagnostic benchmark, not a leaderboard chase: the headline and supporting findings are intentionally falsifiable and the single-eval entrypoint (`eval.py`

-task . . .) ships with fixed seeds and pinned baselines so that improvements can be attributed to method changes rather than evaluation drift.

A.6 Distribution

Release plan. The Tier-1 distilled release is already available for anonymous review at https://anonymous.4open.science/r/VeriBenchWild-60F0; the full dataset (Tier-1 public + Tier-2 gated) will be publicly released for the EMNLP camera-ready, hosted on HuggingFace Hub and Zenodo. Distribution follows the two-tier scheme described below. **Tier 1 (public)** ships JSONL files with `claim_text` either re-hydratable via `tweet_id` (CC-BY-NC-SA-4.0) or replaced by a 50-character truncated preview where re-hydration is not possible; it ships with the hydration script described above. **Tier 2 (gated)** ships verbatim tweet text intact, with access governed by a click-through agreement enforcing X Developer Policy (non-commercial research only, no redistribution of verbatim tweet content outside derivative publications, 7-day takedown SLA, no re-identification of hashed authors). Hosting platform options are HuggingFace Hub gated repo, Zenodo Restricted, or a self-hosted click-through. Distribution begins at EMNLP 2026 camera-ready (anticipated 2026-08); the v4.1.1 snapshot is already frozen with a published SHA-256 manifest.

Licensing (multi-component release). The VERIBENCH-WILD release contains components under multiple licenses. The two licence regimes apply per-tier and are also recorded per-record in the `license` field of every JSON; the `upfd_legacy_only/` sub-directory carries its own dedicated `LICENSE` file so that downstream tooling can split on directory boundary:

- **`upfd_legacy/` tier** ($n = 5{,}147$ total records of which $5{,}086$ carry cascade graphs, $\sim$19 % of the corpus): **CC-BY-SA-4.0**, inherited verbatim from the upstream UPFD release (Dou et al., 2021) to honour ShareAlike on the original PolitiFact / GossipCop cascades.
- **`wild/`, `wild_support/`, `wild_control/` tiers** ($n = 22{,}072$ records, $\sim$81 % of the corpus): **CC-BY-NC-SA-4.0** for derived metadata, labels, and side-cars produced by us.

Verbatim tweet text in Tier 2 is additionally governed by the X Developer Agreement §II.A and by a click-through gated-access agreement that ships with the Tier-2 release. Other upstream licences are honoured: Community-Notes content (X open-data terms) and MBFC factuality buckets (Idiap research-use mirror; only domain-level buckets are redistributed). No export controls or other regulatory restrictions are known to apply; no personal data of EU residents is collected beyond what is already public on X, and right-to-erasure is handled via the takedown channel.

A.7 Maintenance

Support. The dataset is hosted and maintained by the corresponding author (*[anonymous author]*) and the VERIBENCH-WILD team at *[anonymous institution]*. The institutional support email, HuggingFace dataset repo, Zenodo record, and the gated-access agreement will be added to this section on publication (currently withheld to preserve double-blind review). For now, the public Tier-1 distilled release is available at https://anonymous.4open.science/r/VeriBenchWild-60F0.

Errata, updates, retention. The release ships an errata file covering the two known issues in the current snapshot (pre-/post-dedup count discrepancy and the 21-record `user'<hash>` handle-hash artifact, both documented above in App. A.2); future errata append to that file and bump the patch version (next: v4.1.2). Quarterly snapshots will appear while support is active, and **we commit to maintaining the dataset for at least two years post-publication (through 2028-08)**; after that, the v4.1.1 snapshot remains archived on Zenodo with its DOI and maintenance is best-effort. We do not impose a fixed retention horizon: the Zenodo archive of any locked snapshot is preserved per Zenodo's permanence policy, and superseded snapshots are not removed retroactively because reproducibility of already-published results would otherwise be broken. Older versions remain hosted on Zenodo (DOI per snapshot); active maintenance is on the latest snapshot only.

Contributions. Bug reports for mis-flagged AI-positive labels and typology corrections are accepted at the same contact address and incorporated into the next snapshot's audit trail. The build pipeline scripts are released so downstream researchers can re-run against newer Community Notes snapshots; we welcome merge-back of expanded language coverage or improved AI-flag

regex.

A.8 Compute and Carbon Footprint

Full benchmark reproduction (all five evaluation tasks $\times$ five seeds where applicable, plus the T-Match diagnostic and the four headline + one supporting finding) consumes approximately **120 H100-80GB GPU-hours** on the institutional cluster. At a measured node power of $\sim$600 W including host overhead and the cluster's reported PUE, this is roughly **72 kWh** and $\approx$**29 kg CO_2-eq** at the local grid mix (see §7). The dataset-build pipeline itself (regex, hydration, MBFC join, evidence enrichment, clustering, counterfactual generation) accounts for an additional $\sim$15 GPU-hours, dominated by Mistral-7B axis-flip verification.

B Annotation Guidelines and Per-Axis Agreement

This appendix documents the two human-annotation studies that produced the inter-annotator agreement numbers quoted in the main text, together with the instructions, interface, and disagreement patterns that those numbers summarise. Two separate packets were run:

- **Packet 1** (AI-flag gold validation): 4 PhD-level annotators, 190 records, single binary axis (*was this claim AI-generated?*). Reported Fleiss $\kappa = 0.94$.
- **Packet 2** (six-axis typology): 3 trained annotators, 50 records, six non-exclusive misleadingness axes. Three-rater mean Fleiss $\kappa = 0.47$ (range 0.08 on MISSINGIMPORTANTCONTEXT to 0.89 on MANIPULATEDMEDIA), reported below per axis.

The Packet 2 numbers are deliberately reported at their true value rather than averaged into the larger Packet 1 κ. The per-axis disagreements are themselves informative about where the six-axis schema is ambiguous and where it is not (see Tables 10 and 11, plus the worked disagreements in §B.4).

B.1 Recruitment and Compensation

Packet 1. Four PhD-level annotators (anonymized as R1–R4) were recruited through the authors' research-group network. All four hold or are completing doctoral training in NLP, ML, or an adjacent computational discipline, all read English fluently, and all had prior exposure to misinformation, fact-checking, or content-moderation research before this study.

Packet 2. Three annotators were recruited from the same pool, with the further requirement of prior exposure to the Community Notes platform so they could read CN rationales as a primary signal. All three (R5, R3, R2) returned their judgments before the submission deadline. R3 and R2 also participated in Packet 1; R5 is a Packet-2-only annotator recruited to reduce overlap bias.

Compensation. Compensation details are in the Ethics statement (§7). In brief: the four PhD-level raters who produced the Packet 1 and Packet 2 gold judgments are either co-authors of this paper or acknowledged contributors who benefit academically from the release; no paid crowdworkers were used. Upstream Community Notes raters whose rationales appear in the corpus are volunteer X contributors whose compensation is governed by X and is outside the scope of this release.

B.2 Annotation Interface

Both packets used the same self-contained single-file HTML annotation tool (`annotate_t1_typology.html` for Packet 2; a CSV-based variant for Packet 1). The Packet 2 tool runs entirely in the annotator's browser with no network calls, server, or installation: opening the file in any modern browser is sufficient.

For each record the tool renders the original tweet text, the Community Notes rationale, the platform language tag, and a row of six axis controls each offering YES, NO, and SKIP. A *Download results JSON* button is enabled only after all 50 records have a non-empty value on every axis. Progress is auto-saved to `localStorage` after every click, so partial sessions survive browser restarts on the same machine. The full source is released alongside the dataset at `data/wild/annotation/t1_typology_kappa_v1/annotate_t1_typology.html` so that reviewers can reproduce the exact rater experience.

B.3 Instructions Provided to Annotators

The verbatim instructions document gave annotators four kinds of guidance, which we paraphrase here in full:

Axis definitions. Each of the six axes was introduced with a one-line plain-English definition:

- **Factual Error.** The tweet asserts a verifiable false statement of fact.
- **Manipulated / AI-Generated Media.** An image, video, or audio attached to the tweet is fake, AI-generated, edited, or misattributed to a wrong context.
- **Outdated Information.** The claim was true at some earlier time but is no longer true, and is shared without that temporal context.
- **Missing Important Context.** The claim is technically true but omits context that materially changes its meaning.
- **Unverified Claim As Fact.** A rumour, allegation, or unsubstantiated claim is presented as established fact.
- **Satire Passed Off As Serious.** Content originally satirical or comedic is being shared as if literally true.

Annotators were explicitly told that multiple axes can be YES on the same record — the axes are not mutually exclusive — and that the Community Notes rationale should be read as the primary signal for *why* the post was flagged.

Skip rule. Annotators were instructed to distinguish NO ("read it, confident the axis does not apply") from SKIP ("cannot tell — ambiguous, insufficient context, or language barrier"). Skip was to be used sparingly. For non-English tweets, annotators were told to rely on the note text where possible and to fall back to SKIP only when both tweet and note were inaccessible.

Independence. Raters were forbidden from communicating about specific records during annotation and were not shown each other's judgments at any point. The aggregator script (`compute_kappa.py`) was run only after all raters had submitted their JSON files.

Recovery. Auto-save behaviour and the *Reset* button were documented in writing so that annotators were not forced to complete the packet in a single sitting.

B.4 Worked Disagreement Examples

We sampled five Packet 2 records on which at least two of the three raters disagreed on at least one axis, and discuss each below. We pick these from `sample_50.jsonl` because they show, concretely, where the six-axis schema invites principled disagreement rather than rater error. All author handles in the corpus are already replaced with length-16 hex hashes; we have re-verified that the quoted passages contain no remaining real handles or identifying tokens.

(1) `t1k-013` — tier WILD, English.

> *Tweet:* "I will be spreading very soon." [image attachment]
> *Note:* "Community Notes is a feature used to address and correct misinformation, not spread it. Also, the second image is AI generated and depicts LeBron James, not the author of the Tweet."

Disagreements. On *Factual Error*, `R5` said NO, `R3` said YES. On *Outdated Information*, `R5` said YES, `R3` said NO.

Why this is reasonable. The tweet itself makes almost no factual assertion — the falsehood is carried entirely by the AI-generated portrait implying LeBron James authored the post. One rater read the implicit-authorship claim as a *factual* falsehood; the other read it as a *media* falsehood already covered by the Manipulated Media axis. The Outdated tag is an arguable misclick on the other rater's part, but it tracks a real intuition that some Community-Notes rationales hinge on whether information was once true.

(2) `t1k-014` — tier WILD, English.

> *Tweet:* "Sweden is lovely this time of year. I'm so amped about this hike!"
> *Note:* "This is an AI generated image of Iron Hands [from] Pokemon Scarlet and Violet: https://www.pokemon.com/uk/pokedex/iron-hands"

Disagreements. On *Factual Error*, `R5` said NO, `R3` said YES. On *Missing Important Context*, `R5` said YES, `R3` said NO.

Why this is reasonable. The text alone is literally true (Sweden is in fact pleasant in summer); the misleadingness is purely visual. One rater treated "passing off a Pokemon screenshot as a hiking photo" as a factual claim that fails; the other restricted *Factual Error* to assertions made in the prose and routed this case through *Manipulated Media* plus *Missing Context*. Both readings are defensible and the schema does not currently arbitrate between them.

(3) `t1k-019` — tier WILD, English.

> *Tweet:* "As Klaus Schwab is hospitalised, we have to ask the question; Did he follow his own advice with the vaccine, or did he exercise bodily autonomy and refuse it?"
> *Note:* "There is no verifiable sources to substantiate this claim. [...] OP includes an AI generated

image that purports to show the aforementioned Klaus Schwab which misleads through illustrative framing of an already false statement of fact by a UK Political party."

Disagreement. On *Missing Important Context*, R5 said YES, R3 said NO.

Why this is reasonable. Both raters correctly flagged Manipulated Media (the AI image) and the Unverified Claim. They split on whether the rhetorical framing ("we have to ask the question") adds a separate Missing-Context fault on top of the unverified core. One rater treated the false premise itself as carrying the entire problem; the other treated the rhetorical-question framing as an independent context omission. Discriminating these is genuinely difficult and is a known limitation of multi-axis fault taxonomies.

(4) `t1k-021` — tier WILD_CONTROL, English.

Tweet: "IShowSpeed jumped over 2 cars" [video attachment]
Note: "This is obviously fake. No human can stay in the air for that long. No news source should be reporting it as fact."

Disagreements. On *Factual Error*, R5 said NO, R3 said YES. On *Satire*, R5 said YES, R3 said NO.

Why this is reasonable. The video is clearly edited and the assertion is physically impossible. One rater categorised this as *satire passed off as serious* (a stunt video that some viewers nonetheless took at face value, per the note); the other categorised it as a straight *Factual Error*. The schema distinction between *Satire* and *Manipulated Media* + *Factual Error* for impossible-but-amusing clips is one of the schema's softest seams, and this record sits exactly on it.

(5) `t1k-011` — tier WILD_CONTROL, English.

Tweet: "Trump has been inciting actual terroristic threats in Ohio and quadrupled down despite being fact checked into oblivion about the Haitian community there…and your first instinct is to blame democrats when somebody tries to take another shot at him lol."
Note: "Trump repeated things he saw on 'TV'. There is no evidence of Trump making terroristic threats of violence."

Disagreements. On *Factual Error*, R5 said NO, R3 said YES. On *Missing Important Context*, R5 said YES, R3 said NO.

Why this is reasonable. The disputed core ("terroristic threats") is a legal-evidentiary characterisation: the note says no such threats are documented. One rater treated the characterisation as flatly false (*Factual Error*); the other treated it as a tendentious framing of a real event (*Missing Context*). The CN platform itself labelled both Factual Error and Missing Context as true on this record, illustrating that the schema permits both readings simultaneously and that the disagreement is between two reasonable single-axis choices rather than between a correct and an incorrect one.

Note on R2. In each example above R2's judgment aligned with either R5 or R3; the R5-vs-R3 pairings were selected because they most clearly illustrate the schema seams. Per-record judgments from all three raters are included in the released annotation files.

B.5 Per-Axis Agreement Table

Tables 10 and 11 triangulate where the six-axis typology is reliable from three edge points. Table 10 combines three per-axis signals: the 3-rater Fleiss κ (how much three independent annotators agree on each axis), the mean rater–gold Cohen κ (how well each rater aligns with the Community-Notes platform label, averaged across the three pairings), and the headline T1 macro-F1 (how well a trained classifier scores per axis, 5-seed). Table 11 then decomposes the model-vs-human alignment: each cell is the Cohen κ between the 5-seed soft-voted RoBERTa T1 prediction and one of {R5, R3, R2, gold}. Read the two tables together as: axes with high human-human agreement are also the ones the classifier captures best (*ManipulatedMedia*), and axes where humans diverge are also where the classifier struggles (*Satire*, *MissingContext*).

B.6 Caveats on the Agreement Numbers

We close with three caveats that reviewers should weigh when reading the agreement numbers.

Packet 2 is a complete three-rater study. All three annotators (R5, R3, R2) returned their 50-record judgments before the submission deadline, and the per-axis κ values in Table 10 and in the main-text Table 1 are the three-rater Fleiss κ from the final aggregator run on all three rater files. The headline $\bar{\kappa} = 0.47$ is the unweighted mean across the six axes of those three-rater Fleiss values. No part of the agreement story is deferred to camera-ready; the numbers here are final for submission.

Packet 1's $\kappa = 0.94$ is partly keyword-anchored. The binary AI-flag axis benefits from the fact that

Axis	Fleiss κ (3-rater)	Mean Cohen κ rater–gold	T1 F1 (5-seed)
Factual Error	0.401	0.156	0.651 ± 0.022
Manipulated / AI-Generated Media	**0.891**	**0.849**	$\mathbf{0.808 \pm 0.013}$
Outdated Information	0.464	0.310	0.351 ± 0.019
Missing Important Context	0.080	0.239	0.641 ± 0.026
Unverified Claim As Fact	0.380	0.509	0.452 ± 0.031
Satire (passed off as serious)	0.603	0.210	0.220 ± 0.015
Mean across axes	**0.470**	**0.379**	$\mathbf{0.521 \pm 0.020}$

Table 10: Per-axis agreement (Packet 2, n=50, three raters) plus the headline DeBERTa-v3-large T1 macro-F1 per axis (5-seed, rightmost). The classifier ordering tracks the human-agreement ordering — best on MANIPULATED-MEDIA (F1 = 0.81, κ = 0.89), worst on SATIRE (F1 = 0.22) — a third independent signal (alongside humans and the zero-shot NLI baseline of App. D.7) that Finding 4's classifier–judge gap is schema-driven, not an under-training artefact.

Axis	model–R5	model–R3	model–R2	model–gold	Mean human-side
Factual Error	0.011	0.261	−0.073	0.376	0.066
Manipulated / AI-Generated Media	0.754	0.876	0.836	0.750	0.822
Outdated Information	0.075	0.050	0.128	0.237	0.084
Missing Important Context	−0.019	0.089	0.258	0.151	0.109
Unverified Claim As Fact	0.331	0.203	0.362	0.307	0.299
Satire (passed off as serious)	0.073	−0.003	0.071	0.186	0.047

Table 11: Model-vs-human triangulation across all three raters. Cohen κ between the 5-seed soft-voted RoBERTa T1 prediction and each reference. The model tracks humans well on the visual-evidence axis (*Manipulated Media*, $\bar{\kappa} = 0.82$) and poorly on the inference-heavy axes (*Missing Context* 0.11, *Satire* 0.05), mirroring the human-vs-human ordering in Table 10.

Community-Notes authors very often type the literal string "AI generated" or "AI-generated" into the rationale itself. Raters read the note before answering, which means a non-trivial fraction of the headline agreement is anchored on a shared explicit cue rather than on independent forensic judgment. We document this rather than hide it: the AI-flag κ exceeds previously published fact-verification corpora (AVeriTeC 0.62, FEVER 0.68, AVerImaTeC 0.74) but the comparison is not a strict apples-to-apples one because those corpora annotate multi-class verdicts rather than a single keyword-rich binary.

The $\kappa = 0.08$ on Missing Context is real disagreement, not a bug. *Missing Important Context* is the lowest-agreement axis in the three-rater run, with Fleiss $\kappa = 0.08$ (essentially chance). The MissingContext axis's near-zero κ reflects genuine schema ambiguity — annotators disagreed not on facts but on what counts as "missing context". This is a real finding about typology design, not annotation error. We re-verified the aggregator output by hand: across the three raters, one systematically treated any unmentioned-but-relevant fact as "missing context", another restricted the label to cases where the omission flips the apparent meaning, and the third took an intermediate position. The platform itself uses Missing Context as a soft category, and the worked examples (§B.4, especially `t1k-019` and `t1k-011`) illustrate the conceptual seam that produces the disagreement. We report this near-chance value rather than collapsing the axis because the disagreement is a finding about the schema, not a flaw in the measurement.

C Per-Task Hyperparameters and Compute

Per-task hyperparameters are in Tables 12, 13, 14, 15, 16, and 17.

For every supervised baseline, the following six tables list the exact hyperparameters logged by the training harness. Floor baselines (`random`, `majority`, `length_logreg`, `tfidf_logreg`, `tfidf_topic_overlap`, `length_longer`, `author_history_rate`) require no training and are omitted from the per-baseline tables but counted in the compute budget below. All runs use 5 seeds ($\{7, 13, 42, 43, 44\}$) on H100 80GB nodes except CPU-only baselines (LightGBM, logistic regression).

Shared settings. Unless otherwise noted in the per-task tables below, all supervised transformer baselines use: AdamW (Loshchilov and Hutter,

2019) with $\beta = (0.9, 0.999)$ and weight decay 0.01; a linear LR schedule with warmup (warmup ratio specified per task); label smoothing 0; and best-validation checkpoint selection on the task primary metric. Per-task deviations from these defaults (cosine schedule, different warmup ratio, focal/R-Drop variants, etc.) are flagged in the corresponding caption or the row's Notes column. The two zero-shot machine-text detectors in T2-HP continue a line of prior work that includes Grover (Zellers et al., 2019) and DetectGPT (Mitchell et al., 2023).

Table 12 lists the per-baseline T1 typology hyperparameters. All trainable rows use multi-label binary-cross-entropy with positive-class reweighting; checkpoint selection on validation macro-F1.

Table 13 gives the T2-HP human/AI-provenance hyperparameters. The two trainable rows pose AI-attribution as binary classification with class-balanced loss; the bottom rows are gradient-free zero-shot detectors.

Table 14 reports the T0 counterfactual-generation hyperparameters. Mistral-7B is fine-tuned via QLoRA (4-bit NF4); Qwen-2.5-0.5B is fully fine-tuned because it fits a single H100.

Table 15 lists the T-Prop cascade-classifier hyperparameters. All GNNs use balanced class weights and grad-clipping; the LightGBM row uses eight hand-crafted cascade-shape features in lieu of message passing.

Table 16 lists the T-Match pair-discrimination hyperparameters. Trainable pair classifiers use a $\text{dropout} \rightarrow \text{linear}(d \rightarrow 2)$ head; the zero-shot judges replicate the App. E prompts.

Table 17 lists the T-Author trajectory-forecasting hyperparameters. Sequence models consume up to five prior Community-Notes flags per author, with temporal cut-over preventing leakage.

Compute summary. Aggregate wall-clock budget across the entire benchmark (5 seeds $\times$ all baselines, plus LLM-judge passes and the diagnostic ablations of §6) is approximately $\sim$120 GPU-hours on H100 80GB. Per-baseline averages are derived from `train_metrics.jsonl` step counts and measured throughput on single H100s (single-GPU sole-tenant); CPU-only baselines are excluded. Table 18 summarises the breakdown.

Table 18: Compute-budget breakdown on H100 80GB. Diagnostic ablations cover the era-drift probe, per-language slicing, axis-causality test, cross-task transfer, cascade-curation analysis, and LIAR/AVeriTeC cross-benchmark probes. Per-seed wall-time is not written into `eval_results.json` for most baselines; the figures above should be read as $\pm 30\%$ approximations.

Component	GPU-hours
T1 (5 models $\times$ 5 seeds)	$\sim$10
T2-HP (2 trained $\times$ 5 seeds + 2 zero-shot)	$\sim$4
T0 (2 SUTs $\times$ 5 seeds + LLM-judge passes)	$\sim$18
T-Prop (3 GNNs $\times$ 5 seeds + LightGBM CPU)	$\sim$2
T-Match (2 trained + 3 judges $\times$ 5 seeds)	$\sim$9
T-Author (3 sequence models $\times$ 5 seeds)	$\sim$2
Diagnostic ablations (§6)	$\sim$70
Total	**$\sim$120**

The Phi-3-medium and Mistral-7B LLM-judge passes account for the majority of T0 and T-Match wall time and use bf16 inference with greedy decoding (cf. Appendix E).

Reproducibility notes.

1. **T1 / `deberta_v3_large`** v1 (focal+R-Drop) is preserved for audit; the variant headlined in §5 is the retuned BCE+pos_weight `deberta_v3_large_retuned` (Table 12).
2. **T0 Qwen-0.5B and T-Prop UPFD-legacy** were trained before per-seed hyperparameter archival was enforced; their hyperparameters are recovered from adjacent-seed configurations and the training-script defaults. T-Prop's UPFD-legacy variant is single-seed and flagged n_{seeds}=1 in the §5 table.
3. **T-Match Llama-3.1-8B judge** produced no sealed predictions in the locked run and is omitted from the headline T-Match table; its prompt (identical to `phi3_medium_judge`) is in Appendix E.
4. **Early-stopping policy varies by task.** Trainable baselines use validation-metric early stopping with patience= 3 epochs by default; T-Prop GNNs use patience= 10 because per-epoch validation is noisier on graph batches, and the T0 generative head uses no early stop (fixed-epoch training, since validation loss does not track the downstream axis-flip rate the head is judged on).
5. **Bootstrap CIs are seed-fixed.** All 95% bootstrap confidence intervals in the headline table and per-seed ladders use B=10,000 paired resamples with NumPy seed 12345; rerunning the released aggregator on the per-seed predic-

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
roberta_base	roberta-base (Liu et al., 2019)	2e-5	16×3	BCE+ pos_w (fp32)	—
deberta_v3_base	deberta-v3-base (He et al., 2023)	2e-5	16×3	BCE+ pos_w (fp32)	—
mdeberta_v3_base	mdeberta-v3-base	2e-5	16×3	BCE+ pos_w (fp32)	—
modernbert_base	ModernBERT-base (Warner et al., 2024)	2e-5	16×3	BCE+ pos_w (fp32)	—
deberta_v3_large_retuned	deberta-v3-large	1e-5	8 (×4 GA) ×3	BCE+ pos_w (bf16)	MaxLen= 512; ES val_ macro_f1, pat=1
legacy_roberta	roberta-base (legacy 5-seed)	2e-5	16×3	BCE+ pos_w (fp32)	recov.; cf. caption

Table 12: **T1 — 6-axis Typology (multi-label, primary=macro-F1).** MaxLen=256 unless noted. All five models use a linear classification head of shape `linear(hidden→6)` on the `[CLS]` token; warmup ratio 0.10 (base and large). Checkpoint selection: best validation macro-F1. `deberta_v3_large_retuned` is the second-pass tune that replaced an initial focal+R-Drop run that collapsed on one seed; only the retuned variant is quoted in §5. `legacy_roberta` is a pre-archival 5-seed RoBERTa run; its hyperparameters were recovered from the training record (cf. App. C item 2).

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
roberta_base	roberta-base	1e-5	16 (×2 GA) ×4	BCE+ pos_w =5.9 (bf16)	ES val_ auprc, pat=2
desklib_deberta_v3 _large	ai-text-detector-v1.01 (Desklib, 2024; Dugan et al., 2024)	1e-5	16 (×2 GA) ×4	BCE+ pos_w =5.9 (bf16)	fallback: deberta-v3-large
tfidf_logreg	tf-idf + logistic regression	—	—	—	floor recipe (no train cfg)
fast_detect_gpt	tiiuae/falcon-7b (Almazrouei et al., 2023) (scoring LM)	—	4 ×0	analytic sampling disc. (bf16)	zero-shot (Bao et al., 2024)
binoculars _zeroshot	falcon-7b (obs) / falcon-7b-instruct (perf)	—	2 ×0	logppl/xppl (fp16)	threshold 0.9015 (Hans et al., 2024)

Table 13: **T2-HP — Human/AI Provenance (binary, primary=AUPRC).** MaxLen=512 for all transformer/LM rows. Trainable baselines (top two rows) use label smoothing 0.01, warmup ratio 0.06, and a $\mathrm{dropout}(0.1) \rightarrow \mathrm{linear}(d \rightarrow 1)$ head; pos_weight $= 5.9$ matches the training-split class ratio. The two zero-shot detectors share the same data slicer but never see a gradient. `tfidf_logreg` is a floor baseline; its config records only `recipe: floor (no training)`.

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
mistral7b _qlora_axis	Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	2e-4	4 (×8 GA) ×3	causal-LM (bf16, 4-bit NF4)	MaxLen= 2048; QLoRA (Dettmers et al., 2023) r=16, α=32, drop=0.05; cosine; warmup =0.03
qwen_05b _sft_v4_1_1	Qwen2.5-0.5B-Instruct (Yang et al., 2024)	2e-5	8 (×2 GA) ×3	causal-LM (bf16)	MaxLen= 1024; full FT; cosine; warmup =0.03; ES val_ loss, pat=3

Table 14: **T0 — Axis-Controlled Counterfactual Generation (primary=LLM-judge macro axis-flip rate).** Mistral-7B uses QLoRA (NF4 double-quant with bf16 compute) on $\{\text{q,k,v,o,gate,up,down}\}_{\text{proj}}$; Qwen-2.5-0.5B is full-FT (model fits on one H100). Both share greedy decoding at test time (do_sample =False, $\text{max_new_tokens} = 192$, repetition penalty 1.05). Checkpoint selection: best validation loss (surrogate); the headline metric is the Mistral-7B LLM-judge axis-flip rate computed post-hoc (Appendix E).

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
graphgps	GPSConv (Rampášek et al., 2022) ×4 (GINE+ Transformer), h=128, 4 heads	3e-4	32 (×2 GA) ×30	CE+ cls_w bal., LS=0.05 (fp32)	ES pat=8, LapPE k=8, RWPE 16
upfd_gatv2	GATv2Conv (Brody et al., 2022) ×2, h=128, 4 heads, drop=0.2	5e-4	64 (×2 GA) ×10	CE+ cls_w bal., LS=0.05 (fp32)	ES pat=4
upfd_ graphsage	SAGEConv (Hamilton et al., 2017) ×2, h=128, mean agg.	1e-3	128 ×50	CE+ cls_w bal. (fp32)	ES pat=5
lightgbm _8features	LightGBM (Ke et al., 2017) (n=500, lr=0.05, leaves=31)	0.05	—	binary_logloss (cpu)	8 hand-crafted cascade feats
upfd_legacy _roberta	roberta-base text-only (re-implementation)	2e-5	16 ×3	BCE (fp32)	MaxLen= 192; legacy single-seed

Table 15: **T-Prop — Propagation Verdict (binary, primary=macro-F1).** All trainable GNNs use grad-clip $\|g\|_2 \leq 1.0$, balanced class weights, and the `upfd_legacy_only/{train,val,test}.jsonl` splits. The LightGBM features are {n_nodes, depth, max_breadth, mean_branching_factor, structural_virality, wiener_index, root_out_degree, time_to_50pct_reach_s}. *Caveat.* `upfd_legacy_roberta` is a legacy single-seed run; its hyperparameters (epochs=3, BS=16, LR=2e-5, max_len=192) were recovered from the training record.

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
`deberta_v3 _large_pair`	deberta-v3-large	1e-5	8 (×4 GA) ×8	CE, LS=0.05 (bf16)	ES val_pair _acc, pat=3
`roberta_base _pair`	roberta-base	2e-5	16 (×2 GA) ×5	CE (fp32)	ES val_pair _acc, pat=3
`nli_deberta _zeroshot`	DeBERTa-v3-base-mnli-fever-anli (Laurer et al., 2023)	—	64 ×0	entail/contra scoring	zero-shot, no grad
`phi3_medium _judge`	Phi-3-medium-4k-instruct (Abdin et al., 2024)	—	8 ×0	greedy A/B decode	MaxLen= 2048; zero-shot LLM-as-judge
`length_longer`	"longer claim" rule	—	—	—	floor (no training)
`tfidf_topic _overlap`	tf-idf cosine to topic	—	—	—	floor (no training)

Table 16: **T-Match — Topic-Controlled Pair Discrimination (primary=pair accuracy).** MaxLen=512 for all trainable and NLI rows unless noted. Trainable pair-classifiers use $\text{dropout} \rightarrow \text{linear}(d \rightarrow 2)$ heads with warmup ratio 0.06. The `phi3_medium_judge` "system" prompt is reproduced verbatim in Appendix E. A 5th zero-shot judge (Llama-3.1-8B-Instruct (Grattafiori et al., 2024)) shares the prompt of `phi3_medium_judge` but is not part of the locked table.

Baseline	Backbone	LR	BS× Ep	Loss / Prec	Notes
`transformer _history`	TransformerEnc (d=256, 4L, 8H, ff=1024)	3e-4	64 ×8	BCE, LS=0.02 (fp32)	SBERT (Reimers and Gurevych, 2019) MiniLM-L6 (Wang et al., 2020) feats; in_dim =389; hist=16
`gru_history`	GRU(d=386, h=256, 1L, uni)	1e-3	64 ×30	BCE (fp32)	SBERT MiniLM-L6 feats; hist=16
`mamba_hawkes`	MambaHawkes (Gu and Dao, 2024; Hawkes, 1971) (d=256, 4L, d_state=16, d_conv=4, expand=2)	3e-4	64 ×12	BCE+ 0.1MSE (int., $\log 1 + \Delta t$) (bf16)	cosine LR, ES pat=4
`logreg_ history_stats`	sklearn LogReg (12 hand-crafted feats)	—	— × 2000 iter	l2, C=1.0, lbfgs, cls_w bal. (cpu)	12 hist+rate features
`author_ history_rate`	last-window AI rate	—	—	—	floor (no training)

Table 17: **T-Author — Author Trajectory Forecasting (primary=ROC-AUC).** Sequence baselines all use max history length 16, the same SBERT `all-MiniLM-L6-v2` text-embedding backbone (frozen), and early-stop on validation ROC-AUC. The Mamba-Hawkes SSM versions logged at the time of the run are `mamba_ssm==2.3.2.post1` and `causal_conv1d==1.6.2.post1`; this is the version pinned in '`RELEASE_VERIBENCH_WILD/environment.yml`.

tions reproduces the reported CIs to the third decimal place.

D Full Per-Seed Baseline Ladder

This appendix provides the complete baseline ladder per task, with 5-seed mean ± std and 95% bootstrap confidence intervals. Rows marked *(det.)* are deterministic at inference (random/majority/length-only/ TF-IDF floors and zero-shot inference passes) and reported on a single seed. Architecture citations are in Appendix C.

D.1 T1 – Six-axis Typology (macro-F1)

Baseline	**mean ± std**	**95% CI**	**seeds**
random	0.1853 ± 0.0073	[0.178, 0.190]	5
majority	0.0000 ± 0.0000	[0.000, 0.000]	5
length_logreg	0.3224 (det.)	—	5
tfidf_logreg	0.4994 (det.)	—	5
deberta_v3_large_retuned	0.5206 ± 0.0178	[0.503, 0.532]	5
modernbert_base	0.5234 ± 0.0066	[0.518, 0.529]	5
mdeberta_v3_base	0.5231 ± 0.0035	[0.520, 0.526]	5
deberta_v3_base	0.5299 ± 0.0018	[0.528, 0.531]	5
legacy_roberta	0.5315 ± 0.0023	[0.530, 0.533]	5
roberta_base (head)	$\mathbf{0.5327 \pm 0.0021}$	[0.531, 0.535]	5

Table 19: Per-seed scores, T1 six-axis typology (macro-F1).

D.2 T2-HP – AI-Attribution (AUPRC)

Baseline	**mean ± std**	**95% CI**	**seeds**
binoculars_zeroshot	0.1815	—	1 (det.)
fast_detect_gpt	0.1997	—	1 (det.)
majority	0.3682	—	5
random	0.3737 ± 0.0097	[0.364, 0.379]	5
roberta_base	0.6030 ± 0.0042	[0.599, 0.607]	5
length_logreg	0.6184 (det.)	—	5
desklib_deberta_v3_large	0.6366 ± 0.0081	[0.629, 0.643]	5
tfidf_logreg (head)	**0.7782** (det.)	—	5

Table 20: Per-seed scores, T2-HP AI-attribution (AUPRC).

D.3 T0 – Counterfactual Generation (LLM-judge axis-flip)

Baseline	**mean ± std**	**95% CI**	**seeds**
qwen_05b_sft (v4.0)	0.6533	—	1
qwen_05b_sft_v4_1_1	0.6995 ± 0.0059	[0.695, 0.706]	4
mistral7b_qlora_axis (head)	$\mathbf{0.7990 \pm 0.0108}$	[0.788, 0.807]	5
Same outputs scored by a cross-family Phi-3-medium-4k-instruct judge (Finding 2):			
qwen_05b_sft_v4_1_1	0.3932 (det.)	[0.370, 0.418]	4
mistral7b_qlora_axis	0.4819 (det.)	[0.458, 0.506]	5

Table 21: Per-seed scores, T0 counterfactual generation (LLM-judge axis-flip rate). The Qwen mean is over 4 seeds $\{7, 13, 43, 44\}$.

D.4 T-Prop – Cascade Propagation (macro-F1)

Baseline	mean ± std	95% CI	seeds
majority	0.3575	—	5
length_logreg	0.4367 (det.)	—	5
random	0.4943 ± 0.0147	[0.481, 0.507]	5
tfidf_logreg	0.8051 (det.)	—	5
upfd_legacy_roberta	0.8107	—	1 (legacy)
lightgbm_8features	0.8706 (det.)	—	5
upfd_gatv2	0.9477 ± 0.0019	[0.946, 0.949]	5
upfd_graphsage	0.9494 ± 0.0063	[0.944, 0.955]	5
graphgps (head)	$\mathbf{0.9683 \pm 0.0015}$	[0.967, 0.970]	5

Table 22: Per-seed scores, T-Prop cascade propagation (macro-F1).

D.5 T-Match – Topic-Controlled Helpfulness (pair-accuracy)

Baseline	mean ± std	95% CI	seeds
tfidf_topic_overlap	0.4552 (det.)	—	5
nli_deberta_zeroshot	0.4690	—	1 (det.)
mistral	0.4799	—	1 (legacy)
random	0.4828 ± 0.0478	[0.441, 0.524]	5
phi3_medium_judge	0.4966	—	1 (det.)
roberta_base_pair	0.5241 ± 0.0315	[0.492, 0.543]	5
deberta_v3_large_pair	0.5366 ± 0.0281	[0.515, 0.563]	5
roberta (legacy)	0.5431	—	1 (legacy)
majority_A	0.5448 (det.)	—	5
length_longer (head)	**0.5793** (det.)	—	5

Table 23: Per-seed scores, T-Match topic-controlled pair discrimination (pair accuracy).

D.6 T-Author – Trajectory Forecasting (ROC-AUC)

Baseline	mean ± std	95% CI	seeds
random	0.4988 ± 0.0086	[0.491, 0.506]	5
majority	0.5000	—	5
length_logreg	0.7335 (det.)	—	5
author_history_rate	0.8048 (det.)	—	5
logreg_history_stats	0.8152 (det.)	—	5
mamba_hawkes	0.8291 ± 0.0009	[0.829, 0.830]	5
gru_history	0.8323 ± 0.0048	[0.828, 0.836]	5
transformer_history (head)	$\mathbf{0.8327 \pm 0.0014}$	[0.832, 0.834]	5

Table 24: Per-seed scores, T-Author trajectory forecasting (ROC-AUC).

Coverage summary. All trainable baselines are reported at 5 seeds. Deterministic floors (random, majority, length-only, TF-IDF) and zero-shot inference passes are reported at a single seed and marked *(det.)*. Per-seed `eval_results.json` files and the aggregation script are included in the release.

D.7 Zero-shot NLI control for Finding 4 (classifier-undertraining attack)

To control for the possibility that the 3% TFR of our trained T1 classifier (vs 48% TFR of the Phi-3 judge, §6.5) reflects insufficient classifier capacity rather than a real schema-difficulty signal, we ran an independent zero-shot NLI baseline (`MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`) on the same axis-flip test set. The NLI baseline uses *no* axis-specific fine-tuning, *no* labels from our training distribution, and a fundamentally different classification paradigm (textual entailment of an axis-hypothesis template).

Operationalization. For each (original, modified) text pair targeting axis a, we score textual entailment of an axis-specific hypothesis template (e.g., for FACTUALERROR: "This text contains a factually incorrect statement."). The zero-shot targeted-axis flip rate (TFR_{NLI}) is the fraction of records where the hypothesis is entailed by the original but *not* entailed by the modified text.

Result. The macro unconditional TFR is **0.083** (8.3%), much closer to T1's $\sim 3\%$ than to Phi-3's $\sim 48\%$ (per-axis breakdown in Table 25). Macro TFR is computed as the unweighted mean of per-axis unconditional TFR across the six typology axes (FACTUALERROR, MANIPULATEDMEDIA, OUTDATEDINFORMATION, MISSINGIMPORTANTCONTEXT, UNVERIFIEDCLAIMASFACT, SATIRE), matching the macro-aggregation used for the Phi-3 judge and T1 classifier metrics in §6.5 for direct comparability. This triangulates the schema-difficulty interpretation: an independent untrained classifier — with no exposure to our training labels, no axis-specific tuning, and a different classification paradigm — also fails to detect axis-flip events that Phi-3 "confirms". Together with the human-κ triangulation in Table 1, the schema-difficulty interpretation is now defended three independent ways (human, trained classifier, untrained NLI).

Axis	n	TFR_{NLI}
MISLEADINGFACTUALERROR	623	0.082
MISLEADINGMANIPULATEDMEDIA	369	0.011
MISLEADINGOUTDATEDINFORMATION	250	0.024
MISLEADINGMISSINGIMPORTANTCONTEXT	728	0.095
MISLEADINGUNVERIFIEDCLAIMASFACT	391	0.087
MISLEADINGSATIRE	100	0.200
Macro mean	—	**0.083**
Reference (not NLI; from §6.5):		
T1 classifier (5-seed, seed 42 ref.)	2,461	~ 0.03
Phi-3 LLM judge (5-seed mean)	1,263	0.4933

Table 25: Zero-shot NLI control for Finding 4 (§6.5). The NLI baseline (model name in prose above) uses no axis fine-tuning and an entailment-template operationalisation. Macro TFR is closer to T1 ($\sim 3\%$) than to Phi-3 (48%), defeating the T1-under-training attack on Finding 4.

E LLM-Judge Prompt Templates

We use LLM-as-judge in two places in the benchmark: (i) the T0 axis-flip evaluator for axis-controlled counterfactual rewrites, and (ii) the T-Match pair-discrimination zero-shot judges. All four local judges run on H100 GPUs in bf16, with `do_sample=False` (greedy decoding) and `local_files_only=True` so that runs are deterministic and offline. The prompt text below is reproduced verbatim from the released runner scripts; we have not paraphrased. Table 27 summarises the decoding setup across all five judges.

E.1 T0 axis-flip judge — Mistral-7B-Instruct-v0.3 (same family)

Model. `mistralai/Mistral-7B-Instruct-v0.3` (Jiang et al., 2023).
Decoding. greedy (`do_sample=False`, `temperature=1.0`, `top_p=1.0`, `max_new_tokens=40`, `pad_token_id=eos`), `use_cache=True`, bf16 on a single H100.
Tokenizer. Chat template applied via `tok.apply_chat_template` with one user-role message of `SYS` + blank line + `P`, and `add_generation_prompt=True`.
Truncation. `max_length=2048`, `gold_trunc` = `gen_trunc` = 280 characters.

SYSTEM PROMPT

```
You are a precise misinformation typology annotator. Output
↪ only the
requested format.
```

User prompt template. `{short_name}` is one of {factualError, manipulatedMedia, outdated, missingContext, unverified, satire}; `{definition}` is the axis definition from Table 26 below.

USER PROMPT TEMPLATE

```
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>
```

Axis definitions (substituted into the prompt).

Example raw output (one judgment, verbatim from seed 42).

EXAMPLE JUDGMENT (seed 42)

```
record_id   :
↪ vw-control-1716438935512154230_1716488363065364506
axis        : misleadingFactualError
gold_claim  : "I coached minor hockey in Alberta for almost
↪ 20 years. As
   part of my on-going qualifications, I had to take safety
   ↪ training
   that focussed on hockey. Never in those 20 years did I
   ↪ once have
   someone tell me that cardiac arrest was a player risk I
   ↪ needed to be
   aware of. NEVER. https:"
generated   : "BREAKING: 17 athletes have suffered sudden
↪ cardiac
   arrests while playing basketball at the Paris Olympics.
   ↪ This is
   unprecedented. The last time anything like this happened
   ↪ was in
   1924, when 15 athletes suffered sudden cardiac arrests.
   ↪ What did
   those athletes have in common? They ..."
raw_output  :
VERDICT: YES
REASON: The rewrite introduces a false claim about the
↪ number of
athletes suffering cardiac arrests at the Paris Olympics.
```

E.2 T0 axis-flip judge — Phi-3-medium-4k-instruct (cross-family)

Model. `microsoft/Phi-3-medium-4k-instruct`.
Decoding. identical to §E.1 (greedy, `max_new_tokens=40`, bf16, batch size 8).
Prompt. The Phi-3 judge uses the same system prompt, user prompt template, axis definitions, parser, and decoding hyperparameters as §E.1; only the loaded model identifier differs. This identity is the design point: holding the prompt fixed isolates the judge-family effect in Finding 2 (same-family judges inflate absolute axis-flip rates by +30 to +53 pp vs. cross-family judges, while the relative Mistral-vs-Qwen ranking gap of ∼8 pp survives the judge swap; see §6.3).

E.3 T-Match pair judge — Phi-3-medium-4k-instruct

Model. `microsoft/Phi-3-medium-4k-instruct`.
Decoding. greedy (`do_sample=False`, `max_new_tokens=8`), bf16, batch size 8, left-padded inputs (causal LM batched generation requires `padding_side="left"`); `max_length=2048`.
Tokenizer. Chat template applied via `tok.apply_chat_template([sys_msg,`

short_name	definition (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 26: The six axis definitions substituted into the user prompt template above.

`user_msg], add_generation_prompt=True)`.

Output parser. regex `r"\b([AB])\b"`; if no A/B letter is emitted the response is counted as `abstain` and recorded as a wrong prediction (mapped to "B" by convention; the abstain rate is reported separately alongside the headline pair accuracy).

SYSTEM PROMPT

```
You are an expert misinformation fact-checker. You will see
↪  a topic and
two short claims about it. Exactly one claim is truthful and
↪  one is
misleading. Decide which claim is the truthful one. Reply
↪  with a single
letter: A or B. Do not explain.
```

USER PROMPT TEMPLATE

```
TOPIC: {topic}

CLAIM A:
"""
{A}
"""

CLAIM B:
"""
{B}
"""

Which claim is truthful? Reply with a single letter (A or B).
```

Example raw output (T-Match seed 42).

EXAMPLE JUDGMENT (seed 42)

```
pair_id     : tmatch-0128
raw_letter  : B
prediction  : B
raw_gen     : "B"
```

E.4 T-Match pair judge — Llama-3.1-8B-Instruct (5th architecture)

Model. `meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024).

Decoding. greedy (`do_sample=False`, `max_new_tokens=8`), bf16, batch size 8, left-padded.

Prompt. Identical to the T-Match Phi-3 judge (§E.3) byte-for-byte; only the loaded model identifier differs.

Status. The Llama-3.1-8B judge did not produce sealed predictions in the locked run: the compute node had no network access and the model snapshot was not present in the local HuggingFace cache. We list the configuration here so that a reviewer who pre-downloads the weights can reproduce the fifth architecture without modification.

E.5 T0 axis-flip judge — Gemini-2.5-Flash (cross-family, frontier)

Model. `gemini-2.5-flash` via the Google `generativelanguage` API (Comanici et al., 2025).

Decoding. greedy (`temperature=0.0`), `max_output_tokens=40`.

Prompt. System prompt, user prompt template, and axis definitions are identical to §E.1 byte-for-byte; we do not reproduce them again here. This identity is the design point of the frontier-judge sanity check: it isolates judge-family effects from prompt effects.

API details. Unlike the four local judges, Gemini runs against the Google API; the API key is loaded from a local environment file at runtime and is never echoed, logged, or committed. Total budget: n=12,305 judgments across 5 Mistral-7B seeds plus 4 Qwen-0.5B seeds, USD $2.18 spend, 0 parse errors.

Result reference. Macro axis-flip rate over the Mistral-7B counterfactuals is 0.266 ± 0.005 ($\kappa_{\text{PHI3,GEMINI}}$=0.39); see Table 7 and §6.3 for the full Finding 2 result.

E.6 Decoding-parameter cheatsheet

The Python sources for every judge ship with the public release bundle.

F Datasets We Audited But Did Not Adopt as Primary

To document curation honesty, we report each fact-verification or misinformation dataset we con-

Use	Model	Sampling	max_new	batch	precision	local cache
T0 judge	Mistral-7B-Instruct-v0.3	greedy	40	8	bf16	yes
T0 cross-family	Phi-3-medium-4k-instruct	greedy	40	8	bf16	yes
T0 frontier	Gemini-2.5-Flash	greedy	40	—	—	API (Google)
T-Match judge	Phi-3-medium-4k-instruct	greedy	8	8	bf16	yes
T-Match 5th	Llama-3.1-8B-Instruct	greedy	8	8	bf16	MISS (run failed)

Table 27: The four local judges run with `do_sample=False` and the model's native chat template; no API calls are required for those runs. The Gemini-2.5-Flash frontier judge (added as a Finding 2 sanity check) runs via the Google API with a prompt that is byte-for-byte identical to the Phi-3 cross-family judge, so the local-vs-API comparison isolates the judge family rather than the prompt. This deliberate mix of local and API judges follows the cross-family-judge protocol of Zheng et al. (2023) adapted to the misinformation-typology setting.

sidered adopting as the primary substrate for VERIBENCH-WILD, and the specific reason we chose not to.

LIAR (Wang, 2017). 12,836 short political claims, six-class verdict. We use LIAR in Finding 3 as a cross-benchmark probe target. We did not adopt it as primary because (i) the claims are single-sentence, lack social-graph propagation context, and lack a separate AI-attribution axis; (ii) our binary-collapsed length-pair probe showed the construction is itself subject to a length shortcut (pair-accuracy 0.554).

AVeriTeC (Schlichtkrull et al., 2023). $4{,}568$ claims with QA-decomposition evidence and a four-class verdict. We use AVeriTeC in Exp A as a negative cross-benchmark probe (length pair-accuracy 0.484, chance). We did not adopt it as primary because (i) its scale is much smaller than the post-LLM Community-Notes stream and (ii) it does not include propagation cascades or AI-attribution labels that our five evaluation tasks (and the T-Match diagnostic) require.

MuMiN (Nielsen and McConville, 2022). A 21-million-node misinformation social network. We considered MuMiN as a propagation-graph backbone but did not adopt it because (i) the underlying tweets require hydration that is now blocked by X API policy, making reproducibility uncertain, and (ii) the temporal coverage stops before the post-LLM era boundary that Finding 1 requires.

FEVER (Thorne et al., 2018) and FEVEROUS (Aly et al., 2021). We did not include FEVER as a sub-corpus because its synthetic claim-generation pipeline (Wikipedia rewrites) does not match the naturalistic Community-Notes prose we target; FEVEROUS adds structured-table evidence but inherits the same synthetic-claim construction. Including either would introduce a third confounded programme axis on top of the pre/post-LLM split (Finding 1).

MultiFC (Augenstein et al., 2019). A real-world multi-domain dataset of 36,534 claims from 26 fact-checking websites. We did not adopt it because (i) the verdict taxonomies are site-specific and would require an additional cross-site reconciliation step that we considered out of scope, and (ii) the dataset is text-only and does not carry propagation or AI-attribution signals.

PUBHEALTH (Kotonya and Toni, 2020). A domain-specific corpus of 11,832 public-health claims with explanations. We did not adopt it as primary because the domain restriction would narrow the audit framing to a single topical axis; we include it here as part of the dataset landscape we explicitly considered.

Synthetic AI-misinformation corpora (e.g., RAID). We considered training T2-HP on synthetic-AI corpora and evaluating on VERIBENCH-WILD, and do include the desklib model trained on RAID as a T2-HP zero-shot baseline. We did not adopt synthetic-AI as the *training* substrate because the post-LLM Community-Notes attribution is the deployment distribution we care about.

G Responsible NLP Research Checklist

We report our answers to the ARR Responsible NLP Research Checklist (retrieved 2026-05-21 from `https://aclrollingreview.org/responsibleNLPresearch/`). Table 28 summarises the answers and the paper location that addresses each item. The two non-trivial items, D4 (IRB exemption) and E1 (AI-assistant disclosure), are expanded in the paragraphs of this appendix.

D4 – IRB rationale. The dataset combines (i) publicly posted Community-Notes-flagged

Item	Question (abbrev.)	Ans.	Where addressed
A1	Limitations described	Y	§Limitations (8 paragraphs)
A2	Risks discussed	Y	§Ethics (§7)
B1	Creators of artifacts cited	Y	Inline first-mention citations; App. F
B2	License / terms discussed	Y	§Ethics; CC-BY-NC-SA, 2-tier; gated-access agreement ships with release
B3	Use consistent with intent	Y	§Ethics *Author consent*; App. F
B4	PII / anonymization steps	Y	§Ethics *Privacy / linkage* (§7); salted SHA-256 hashing
B5	Artifact documentation	Y	§3.1–§3.4 (incl. Table 3); App. A; release datasheet
B6	Statistics / splits reported	Y	§3.2, §3.4; SHA-256 manifest
C1	Parameters, budget, infra	Y	§5 *Compute*; §Ethics *Carbon cost*; App. C
C2	Hyperparameter search	Y	§5 *Protocol*; App. C
C3	Descriptive statistics	Y	5-seed mean±std + 95% bootstrap CI in Table 4; App. D
C4	Existing packages reported	Y	§5; App. E, C; release `pip freeze`
D1	Annotator instructions	Y	App. B (interface description + rubric text)
D2	Recruitment / payment	Y	§Ethics *Annotator compensation*; App. B
D3	Consent obtained	Y	Annotators: explicit form. Tweet authors: X-platform ToS + takedown protocol (§Ethics)
D4	IRB approval / exemption	N	No IRB filed: PhD annotators from authors' academic network + public ToS content; institutional policy exempts. Disclosed in § G.
D5	Annotator demographics	Y	App. B; aggregated to preserve small-pool annotator privacy
E1	AI-assistant disclosure	Y	§ G; AI writing assistant used for prose and presentation; all empirical content author-generated

Table 28: Responsible NLP Research Checklist summary. Y = yes (paper addresses the item), N = no (with rationale below), NA = not applicable. The 18 items follow the ARR May 2026 question set (A1–A2, B1–B6, C1–C4, D1–D5, E1).

tweets and PolitiFact / GossipCop fact-checks released under their original academic-use terms, and (ii) labels from PhD-level researchers recruited from the authors' academic network rather than from a participant pool. Under the host institution's policy on secondary use of public social-media content, this combination is not classified as human-subjects research and we did not file with an ethics review board. We disclose the non-filing here rather than omit it; authors will reconfirm institutional classification before camera-ready, and convert this row to Y with a determination letter if the institutional policy has changed.

E1 – AI assistant disclosure. An AI writing assistant was used for paper-prose drafting, paragraph restructuring, and figure / table layout suggestions. The assistant was *not* used to generate or fabricate experimental numbers, dataset records, annotator labels, or finding interpretations. All empirical results, dataset construction scripts, model selection decisions, finding interpretations, and per-seed numbers in Table 4 and Appendix D are author-generated and have been independently verified against the locked release. The authors retain full intellectual responsibility for the content, consistent with the ACL publications ethics policy on AI assistance. The specific assistant and version will be named in the camera-ready; anonymising the vendor here preserves double-blind review.

Linkage-attack disclosure (cross-reference for B4). The full linkage-attack quantification and the two-tier release mitigation are documented in the Ethics statement and in the supporting diagnostics (§7).