future internet

MDPI

Article

Credibility Context Improves Political Claim Classification on LIAR and Transfers to LIAR2 NEW

Bushra Alkomah * and Frederick Sheldon *

Department of Computer Science, College of Engineering, University of Idaho, Moscow, ID 83844, USA
* Correspondence: alko6681@vandals.uidaho.edu (B.A.); sheldn@uidaho.edu (F.S.)

Abstract

Automated screening of political claims is challenging because many statements are short, underspecified, and labeled with fine-grained truthfulness categories that are difficult to separate using claim text alone. This study examines whether lightweight credibility context, represented by pre-existing speaker fact-checking history counts, can improve transformer-based political claim classification without external evidence retrieval. We evaluate the LIAR benchmark under both its native six-class formulation and a coarser three-class mapping using two strong pretrained encoders, DeBERTa-v3-large and RoBERTa-large. To isolate the effect of credibility context, we vary only one input factor: whether the five LIAR speaker-history counts (pants-fire, false, barely-true, half-true, mostly-true) are appended to the claim and standard metadata as structured text (Hist ON) or omitted (Hist OFF), while keeping the data split, model architecture, training pipeline, and evaluation protocol fixed. All experiments are repeated across five random seeds (7, 13, 42, 123, 2024) and reported as mean ± standard deviation. On LIAR, Hist ON improves macro-F1 and accuracy across both backbones and both label granularities, with the largest gains in the six-class setting where label ambiguity is highest. In the six-class task, macro-F1 increases from 0.3096 to 0.4773 for DeBERTa-v3-large and from 0.3127 to 0.4855 for RoBERTa-large. In the three-class task, the best Hist ON model reaches 0.6241 macro-F1. Because only five seeds are available, the minimum achievable two-sided paired Wilcoxon p-value is 0.0625; therefore, we do not claim conventional $\alpha = 0.05$ statistical significance and instead report paired mean differences, seed-level gains, and complementary prediction-level reliability analyses from saved test predictions. To assess whether the benefit is limited to the original LIAR test split, we further evaluate the LIAR-trained checkpoints directly on the full non-overlapping LIAR2 NEW test split without additional fine-tuning. This LIAR-to-LIAR2 NEW transfer evaluation shows that Hist ON improves macro-F1 over Hist OFF in all four backbone/granularity settings. The best absolute transferred macro-F1 is achieved by the three-class DeBERTa-v3-large setting (0.6462), whereas the largest Hist ON minus Hist OFF gain occurs in the six-class RoBERTa-large setting (+0.1433 macro-F1). These two values describe different quantities: absolute performance and improvement over the corresponding Hist OFF baseline. We frame the task as political claim classification, or screening, rather than evidence-grounded fact verification: the method uses speaker-level credibility priors and does not retrieve external evidence. The results support speaker-history credibility context as a low-cost, model-agnostic signal for improving claim screening, while the LIAR2 NEW findings should be interpreted as related-benchmark robustness rather than universal out-of-domain generalization.

Academic Editors: Sardar Jaf, Basel Barakat and Yongqiang Cheng

Received: 26 May 2026
Revised: 16 June 2026
Accepted: 24 June 2026
Published: 30 June 2026

Keywords: political claim classification; automated fact-checking; credibility context; speaker history; LIAR dataset; LIAR2; transfer evaluation; DeBERTa; RoBERTa; multi-class classification

1. Introduction

Political statements spread quickly through news media and social platforms, and their framing and perceived accuracy can influence public opinion and policy [1,2]. The volume of such content has grown faster than the capacity of human verification: professional fact-checking is careful and evidence-based, but it is expensive and difficult to scale to the daily volume of political claims [3,4]. This gap has motivated a sustained research program in automated claim screening, with the goal of providing rapid, transparent decision support for fact-checkers, journalists, and platform moderators [5–7].

A common formulation treats claim screening as a supervised classification problem: given a short political statement, a model predicts a truthfulness label drawn from a fixed scale [8,9]. Even with strong pretrained language models, this problem remains difficult because political language is rhetorical, context-dependent, and often underspecified. A statement may omit the relevant time frame, baseline for comparison, or measurement definition needed to judge whether it is technically correct [10,11]. Two superficially similar claims can therefore receive different truthfulness labels, and the difficulty increases when the task requires fine-grained, non-binary verdicts [6].

Throughout this paper, we distinguish four related terms to avoid overclaiming. Political claim classification refers to supervised prediction of a truthfulness label from claim text and metadata, without evidence retrieval. Automated fact-checking is the broader pipeline that may include claim detection, evidence retrieval, evidence reasoning, and verdict prediction [6]. Evidence-based verification explicitly compares a claim against retrieved evidence in order to produce a justified verdict. Speaker credibility modeling estimates a speaker-level prior over truthfulness behavior. Our work performs claim classification augmented with speaker credibility context; it does not retrieve external evidence and therefore should not be interpreted as strict evidence-based verification.

The LIAR benchmark [9] has become a widely used testbed for political claim screening. It contains short statements labeled on a six-level truthfulness scale (pants-fire, false, barely-true, half-true, mostly-true, true) and accompanies each claim with structured metadata about the speaker, including party, state, job, topic, context, and five speaker-history counts that summarize the speaker's prior fact-checked statements at the other five labels. Because the official LIAR splits are not speaker-disjoint, some speakers may appear across the train, validation, and test partitions. This matters for the present study because speaker-history features can improve in-distribution classification while also introducing speaker-shortcut risk; we therefore discuss this limitation explicitly in Section 7. These counts can be interpreted as a lightweight credibility prior that captures a speaker's historical truthfulness pattern without requiring new evidence retrieval. This type of behavioral prior is consistent with practical claim triage: when a speaker has a long record of misleading statements, a new claim from that speaker may require closer scrutiny until independently verified [7,9].

This paper investigates how much of the difficulty of LIAR-style truthfulness classification can be reduced by adding such credibility context, and where the improvement appears. We adopt a deliberately controlled comparison between two input configurations that differ only in whether the five LIAR speaker-history counts are appended to the input as structured textual fields. In the Hist OFF setting, the model receives the claim and

standard metadata only. In the Hist ON setting, the same input is augmented with the speaker-history counts. Everything else, including model family, tokenizer, optimization, schedule, data split, and evaluation protocol, is held fixed. We hypothesize that Hist ON improves over Hist OFF, with the largest gains in the six-class setting where adjacent truthfulness labels are most ambiguous.

The practical motivation for this design is twofold. From a research perspective, the controlled comparison helps distinguish improvements from richer linguistic modeling of the claim from improvements caused by speaker-level credibility priors [12]. From an application perspective, speaker-history counts may be available for known public figures even when evidence retrieval is too slow, costly, or incomplete for every incoming claim [13]. If such information improves classification reliably, it can serve as a low-cost augmentation for screening pipelines. At the same time, this signal has important limitations: a model that relies too strongly on speaker-level history may learn shortcuts and may generalize poorly to new speakers, new election cycles, or speakers with sparse prior records [13,14].

The contributions of this work are as follows:

- Controlled credibility-context ablation on LIAR. We isolate the effect of LIAR speaker-history counts by comparing Hist OFF and Hist ON under the same backbones, preprocessing, official splits, training pipeline, and hyperparameters.
- Multi-seed transformer evaluation across two label granularities. We fine-tune DeBERTa-v3-large and RoBERTa-large on both the native six-class LIAR task and a coarser three-class mapping, reporting mean $\pm$ standard deviation across five seeds.
- Related-benchmark transfer evaluation on LIAR2 NEW. We evaluate LIAR-trained checkpoints directly on the non-overlapping LIAR2 NEW test split without additional fine-tuning to test whether the credibility-context benefit persists beyond the original LIAR test set.
- Imbalance-aware and error-focused analysis. We emphasize macro-F1, report confusion matrices and ROC/precision–recall analyses, and add a quantitative failure-mode breakdown covering adjacent-label errors, over-severe errors, over-lenient errors, conservative-bias errors, and directional false/true transitions.

The rest of the paper is organized as follows. Section 2 surveys related work on LIAR-based claim classification, metadata-augmented modeling, evidence-aware verification, and label ambiguity. Section 3 describes the dataset, input construction, models, training pipeline, and evaluation protocol. Section 4 reports results on the original LIAR benchmark, and Section 5 presents the LIAR-to-LIAR2 NEW transfer evaluation. Section 6 discusses the findings, Section 7 summarizes limitations, Section 8 outlines future work, and Section 9 concludes the paper.

2. Related Work

Automated fact-checking is commonly described as a multi-stage pipeline that may include claim detection, evidence retrieval, evidence reasoning, and verdict prediction [5,6]. Within this pipeline, our work targets the claim-level veracity classification stage, in which a short statement and its associated metadata are mapped to a discrete truthfulness label. We organize the related work along five themes.

2.1. Claim-Level Veracity Classification on LIAR

A central benchmark for claim-level veracity classification of political statements is the LIAR dataset [9]. LIAR pairs short real-world political claims with fine-grained truthfulness labels (pants-fire, false, barely-true, half-true, mostly-true, true) and structured metadata. Subsequent work has confirmed that claim text alone is often insufficient because political statements can be short and contextually underspecified, and that surface linguistic and

stylistic cues correlate with deception only weakly and tend to degrade under domain shift [14].

Several LIAR-focused studies explore lightweight, hybrid, and structured-context architectures and analyze why fine-grained performance remains limited. Kumar et al. [15] propose KGFakeNet, a knowledge graph-enhanced fake news detection model that uses LIAR-PLUS statements, justifications, metadata, OpenIE6 triplet extraction, attention-based alignment, and an MLP classifier for six-class veracity prediction. The recurring message across these studies is that fine-grained truthfulness classification on LIAR-style datasets remains difficult primarily because adjacent labels, such as barely-true versus half-true, are not cleanly separable from claim wording alone.

2.2. Transformer Models for Misinformation and Fact-Checking

Pretrained transformer encoders have become standard backbones for misinformation classification in general and for LIAR-style truthfulness prediction in particular, because they substantially improve over CNN-/RNN-based baselines on short-text classification [6,7]. Recent surveys argue that transformer-only classifiers eventually saturate on LIAR-like tasks unless additional context (metadata, retrieved evidence, or external structure) is supplied [12,13]. Our experiments use DeBERTa-v3-large and RoBERTa-large as strong, comparable text encoders and isolate the contribution of one such additional context signal (speaker history counts).

2.3. Metadata and Speaker Credibility Signals

LIAR's design explicitly invites metadata-aware modeling: each example includes speaker name, party, state, job, topic, context, and five speaker history counts. A common direction adds a separate metadata branch alongside the text encoder so that speaker and context features inform the prediction. More broadly, work on credibility-aware misinformation detection argues that speaker-level cues can substantially improve classification accuracy, but warns that such cues can also encourage shortcut learning and may not transfer to unseen speakers [12,13,16]. A second line of work models richer entity- and topic-structured relationships, for example through knowledge-graph style representations layered over text [15].

The shared concern across this literature is that metadata can become a shortcut: it boosts in-distribution accuracy by encoding properties of known speakers, but offers limited robustness when the speaker mix changes. Our Hist ON versus Hist OFF design is intended to expose, rather than hide, that trade-off: because everything else is held fixed, any change in performance is attributable to the inclusion of speaker history counts, and the same comparison can be re-run on speaker-disjoint splits in future work.

2.4. Evidence-Aware and Explanation-Aware Verification

A complementary line of work moves toward evidence-grounded and explainable verification, in which a system explicitly retrieves and reasons over evidence rather than classifying from text alone. Ref. [17] proposes a framework focused on trustworthiness, generalizability, and controllability, and ref. [15] combines structured knowledge-graph evidence with attention-based integration. These approaches are stronger in principle than retrieval-free classification, but they are also computationally heavier and require evidence stores that may be unavailable in real-time screening pipelines. Our work is intentionally on the opposite end of the cost spectrum: we ask how much credibility context (a single structured field) can buy without retrieval.

2.5. Label Ambiguity, LIAR Refinement, and LIAR-PLUS-Style Resources

Truthfulness is inherently subjective when claims are underspecified or evidence is ambiguous. Recent work explicitly studies annotator disagreement and soft-label perspectives for fact-checking [18,19], showing that ambiguity is a real characteristic of the task rather than label noise to be removed. This justifies common LIAR practices of also evaluating a coarser three-class mapping that collapses adjacent fine-grained labels [9]. The LIAR2 dataset [20] extends and refines this line of work by re-annotating and expanding the original LIAR benchmark with professionally verified labels and additional metadata fields; we use its non-overlapping NEW portion as a transfer-evaluation benchmark for our LIAR-trained checkpoints (Section 5).

2.6. Limitations of Prior Work and Positioning of Our Study

Across the directions above, three gaps are especially relevant to LIAR-style claim classification:

- Isolating credibility priors. Many systems change several components at once (architecture, metadata, objectives, retrieval), which makes it hard to attribute improvements to credibility context specifically.
- Shortcut risk from metadata. Speaker and context features can dominate predictions when claims are short, but ablations that explicitly remove credibility priors are rare.
- Ambiguity-aware error analysis. Adjacent truthfulness labels and limited claim context mean it is necessary to look at where errors concentrate, not only at scalar accuracy.

Our work targets these gaps directly. Section 3 fixes the backbone, training procedure, and data split, and varies only whether speaker history is included; Section 4 reports class-wise behavior and confusion matrices; and Section 5 retests the conclusion by transferring the LIAR-trained checkpoints to the non-overlapping NEW portion of LIAR2.

2.7. Comparative Summary

Table 1 summarizes the main LIAR-related research directions.

Table 1. Qualitative comparison of LIAR-related directions and what they contribute. The placement of our own work in the last row emphasizes that our novelty is the controlled comparison, not a new architecture.

Work	Setting	Signals Used	Takeaway/Limitation
[9]	Claim-level veracity benchmark	Claim text + rich metadata	Establishes fine-grained classification; exposes label ambiguity and metadata shortcut risk.
[14]	Cross-domain deception/misinformation analysis	Linguistic/stylistic cues across domains	Style cues degrade under domain shift and do not substitute for evidence.
[15]	Knowledge-graph enhanced detection	Text + structured relations/metadata	Structured context helps; must guard against shortcuts and weak transfer.
[17]	Explainable, generalizable detection framework	Text + structured reasoning/explanations	Promotes interpretability; emphasizes generalization and controllability.
[18]	Ambiguity-aware fact-checking with evidence	Claim + evidence + soft labels	Shows ambiguity is intrinsic; supports soft-label and reformulation strategies.
[6]	Survey of automated fact-checking	Pipeline view (retrieval, reasoning, verdict)	Organizes task definitions and pitfalls (robustness, explainability, evaluation validity).
[20]	Dataset refinement and extension (LIAR2)	Quality-focused relabeling + extra metadata	Improves evaluation reliability; supports robustness studies beyond the original LIAR artifacts.
[21]	Multimodal deception dataset (DECEPTIcON)	Text + audio + video (PolitiFact labels)	Highlights multimodal cues and dataset variability beyond text-only LIAR.
Ours	Controlled credibility-prior study	LIAR main evaluation; LIAR-to-LIAR2 NEW transfer evaluation	Isolates speaker-history effects under a fixed pipeline, but remains limited by non-speaker-disjoint LIAR splits, related-benchmark transfer only, and unresolved speaker-shortcut risk.

3. Methodology

This section specifies the dataset, input construction, models, training setup, and evaluation protocol with enough detail for an external researcher to reproduce the experiments. The complete pipeline is summarized in Figure 1.

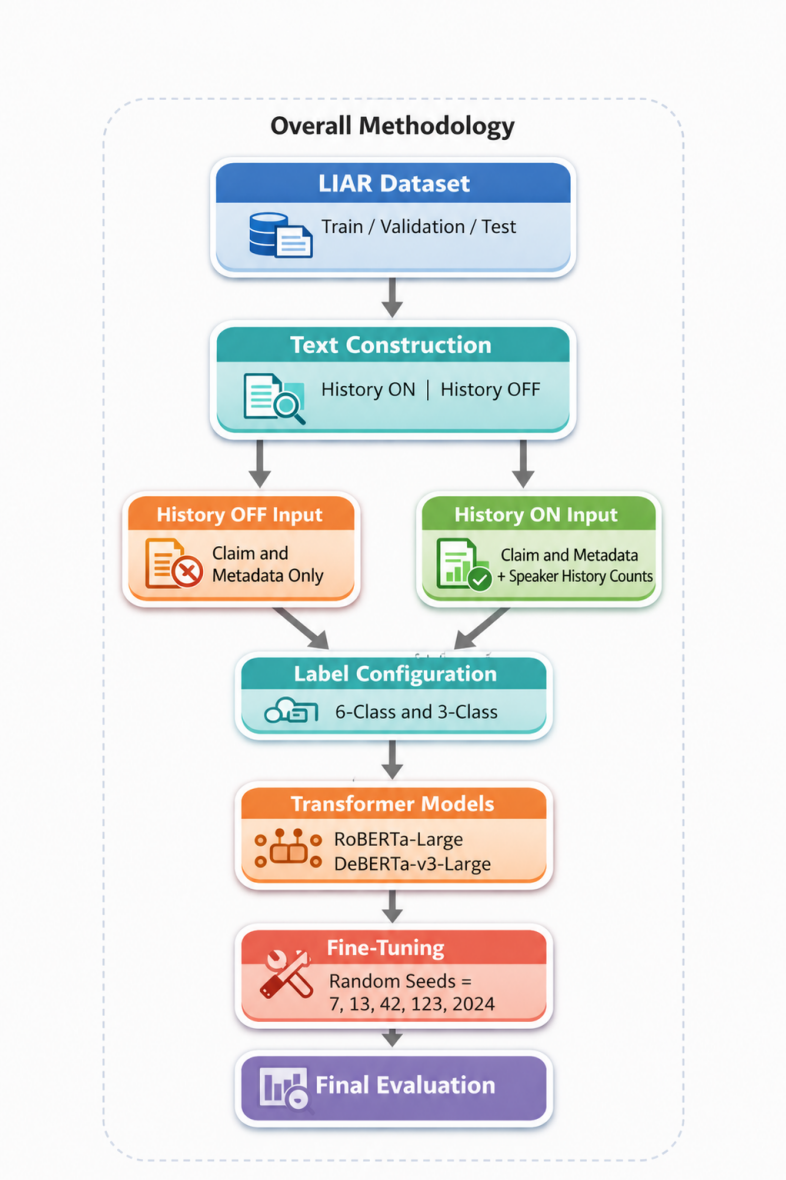


Figure 1. Overall methodology. The pipeline shows the LIAR train/valid/test splits, History ON and History OFF input construction, label-space configuration (6-class versus 3-class), transformer-based fine-tuning of DeBERTa-v3-large and RoBERTa-large, and final evaluation. All configurations are run under the same multi-seed protocol with seeds 7, 13, 42, 123, and 2024; the figure depicts the shared pipeline that is repeated identically for every seed.

3.1. Dataset Description

We conduct the main experiments on the LIAR dataset [9], which contains short political statements annotated by professional fact-checkers into six truthfulness categories: pants-fire, false, barely-true, half-true, mostly-true, and true. Each statement is accompanied by structured metadata: subject, speaker name, speaker job title, state, political party, free-text context, and five speaker history counts (pants-fire, false, barely-true, half-true, mostly-true) summarizing the speaker's prior fact-checked statements. We use the

official train.tsv, valid.tsv, and test.tsv splits supplied with LIAR and do not resample or merge them.

LIAR is imbalanced across categories, particularly at the extreme truthfulness labels, which motivates reporting macro-averaged metrics in addition to accuracy.

In addition to the original six-class formulation, we evaluate a three-class variant in which the six labels are mapped as follows: false = {pants-fire, false, barely-true}, half = {half-true}, and true = {mostly-true, true}. This mapping is a common simplification in the LIAR literature and lets us check whether the trends observed at fine granularity persist when the label space is coarsened.

3.2. Dataset Preprocessing and Formatting

Preprocessing is applied identically to both Hist ON and Hist OFF settings, so that the only difference between the two is whether speaker history counts appear in the constructed input string. The pipeline is sketched in Figure 2.

Step 1: Loading and basic cleaning. Each TSV split is read with the standard LIAR column ordering (label, statement, subject, speaker, job, state, party, history counts, context). All string fields are stripped of leading/trailing whitespace; missing values are replaced with empty strings; labels are lowercased.

Step 2: Label filtering and task setup. Examples whose label is not in the active label set are removed and the index is reset:

- Six-class task: {pants-fire, false, barely-true, half-true, mostly-true, true}.
- Three-class task: The six labels are mapped to {false, half, true} as above.

Step 3: Input text assembly (Hist OFF versus Hist ON). For each example, we build a single text sequence by concatenating the available fields with explicit prefixes and a pipe (|) delimiter:

```
claim: <statement> | subject: <subject> | speaker: <speaker> | job: <job>
| state: <state> | party: <party> | context: <context>
```

In the Hist ON setting, we additionally append a structured `speaker_history` field of the form

```
speaker_history: barely_true_count: <n> | false_count: <n> | half_true_count:
<n> | mostly_true_count: <n> | pants_on_fire_count: <n>
```

For example, an anonymized Hist ON input may take the form

```
claim: The state added 50,000 new jobs last year | subject: economy
| speaker: speaker_A | job: governor | state: state_X
| party: party_Y | context: campaign speech
| speaker_history: barely_true_count: 2 | false_count: 4
| half_true_count: 3 | mostly_true_count: 5
| pants_on_fire_count: 1
```

The corresponding Hist OFF input contains the same claim and standard metadata fields but omits the `speaker_history` segment. Each count is the speaker's prior history count for that label, taken directly from LIAR's history columns.

Step 4: Integer label encoding. The 6-class labels are encoded as integers 0–5 in the order: pants-fire, false, barely-true, half-true, mostly-true, and true; the 3-class labels are encoded as false = 0, half = 1, and true = 2.

Step 5: Tokenization and batching. The constructed text is tokenized with the corresponding pretrained tokenizer of each backbone (RoBERTa-large or DeBERTa-v3-large) with truncation to a fixed maximum sequence length of 192 tokens. Dynamic padding within each batch is used (`DataCollatorWithPadding`) so that padding is bounded by the longest sequence in the batch.

Step 6: Handling class imbalance. Class weights are computed from the training-split label distribution using inverse-frequency weighting and applied through a weighted cross-entropy loss; the procedure is described in Section 3.5.

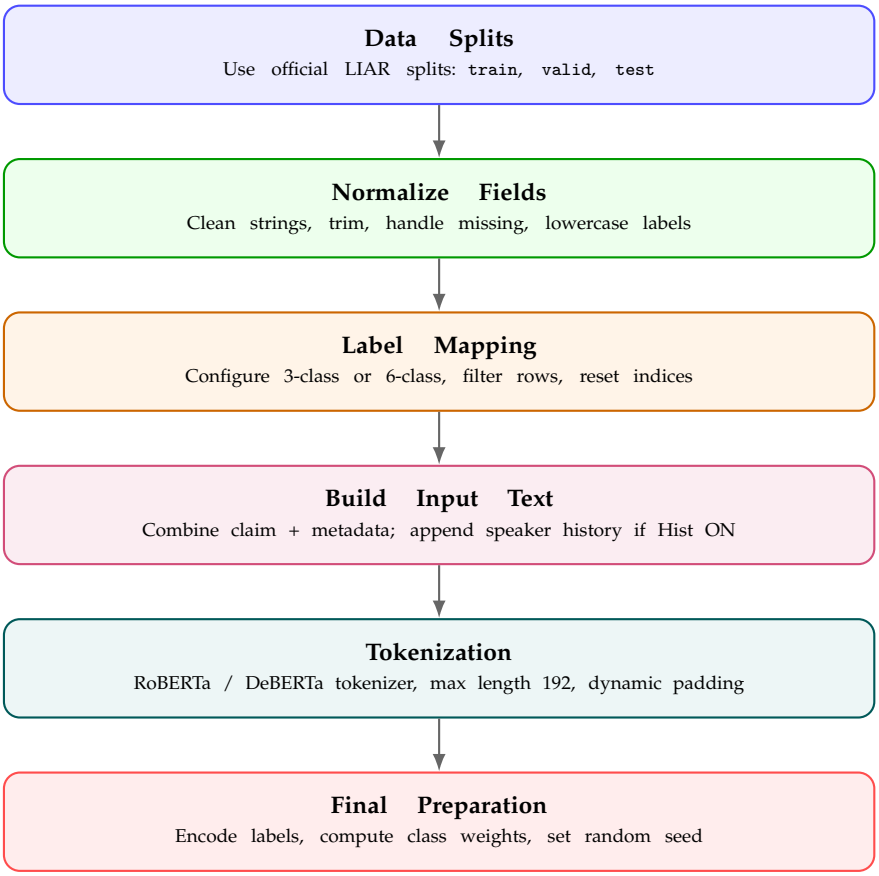


Figure 2. Preprocessing pipeline for the LIAR dataset prior to fine-tuning. The Hist OFF versus Hist ON difference enters at the "Build Input Text" step.

3.3. Hist ON Versus Hist OFF Input Construction

The central manipulation in this study is the binary inclusion of speaker history counts in the input. In the History OFF setting, the model sees only the claim and standard metadata fields. In the History ON setting, the same input is augmented with the five LIAR speaker history counts as a structured textual field, as shown in Step 3 above.

LIAR constructs these history counts from statements that strictly predate the target claim and explicitly excludes the current claim from its own counts, so the count fields do not leak the target label into the input. Both Hist OFF and Hist ON inputs are represented as a single sequence of text with explicit field delimiters, which makes them directly compatible with any standard transformer encoder and means that the only thing the model sees differently between the two settings is the presence or absence of the speaker history field.

We deliberately integrate the speaker-history counts by appending them as structured text, rather than through a dedicated metadata encoder or a learned fusion module. We do not present this as an architectural contribution. The reason for choosing the simplest possible integration is methodological: it is reproducible (no extra components to tune), low-cost (no additional parameters beyond the tokenized text), and model-agnostic (it works with any text encoder without modification), which makes it a clean baseline for measuring the value of the credibility signal itself. A more expressive integration (a separate metadata encoder, or attention-based/gated fusion) might extract more from the same

counts; we did not run those variants, so we do not claim them, and we list them only as a planned ablation in future work (Section 8).

3.4. Model Architectures

We fine-tune two strong pretrained transformer encoders: microsoft/deberta-v3-large and roberta-large, in both cases by attaching a task-specific linear classification head over the pooled sentence representation. The head has six output neurons in the six-class task and three in the three-class task; all model parameters are updated during fine-tuning.

3.5. Training Procedure

All experiments are repeated across five random seeds $\{7, 13, 42, 123, 2024\}$ to account for stochastic variability in transformer fine-tuning. For each combination of (backbone $\in$ {DeBERTa-v3-large, RoBERTa-large}) $\times$ (granularity $\in$ {6-class, 3-class}) $\times$ (input $\in$ {Hist OFF, Hist ON}) $\times$ (seed), we run a separate fine-tuning job.

Training uses the AdamW optimizer with learning rate 1×10^{-5}, weight decay 0.01, a cosine schedule with linear warm-up over the first 6% of training steps, gradient clipping at max norm 1.0, and label smoothing of 0.05. The per-device train batch size is 8 with 4 gradient-accumulation steps, giving an effective batch size of 32; per-device evaluation batch size is at least 32. Maximum number of epochs is 20, with early stopping triggered by validation macro-F1 (best so far over the run) and a patience of 3 evaluation rounds; the best-by-validation-macro-F1 checkpoint is saved and reloaded for test-set evaluation. Mixed-precision training (bf16 when supported by the GPU, fp16 otherwise) is used to reduce memory and time cost. Tokenization uses a maximum sequence length of 192 with dynamic padding.

These hyperparameters are identical for Hist OFF and Hist ON within each backbone–granularity combination, so any test-time difference between the two settings is attributable to the input representation rather than to tuning differences.

Class weights. Class weights are computed from the training split using

$$w_k = \frac{N}{K \cdot n_k},$$

where N is the number of training examples, K is the number of classes, and n_k is the number of training examples in class k. The weights are then renormalized so that their mean equals one, and applied as the `weight` argument of `CrossEntropyLoss`. Class weights are computed separately for each task configuration after label filtering and mapping. Thus, the six-class experiments use weights estimated from the six-class training distribution, while the three-class experiments use weights recomputed from the mapped three-class training distribution. The weights are not shared across label granularities. This reduces the optimizer's bias toward majority labels while preserving overall loss scale; it is paired with the label-smoothing factor noted above for calibration and stability.

Reproducibility controls. We fix Python, NumPy, and PyTorch seeds at the start of every run; pass the same seed as `seed` and `data_seed` to the `TrainingArguments`; set `cudnn.deterministic = True` and disable `cudnn.benchmark`; and disable Hugging Face tokenizer parallelism to remove a source of nondeterminism. The released pipeline additionally exports an `environment.json` with Python, PyTorch, CUDA, and key package versions, plus dataset file checksums, so that runs can be matched to a specific software stack.

3.6. Evaluation Metrics

We evaluate with accuracy and macro-averaged F1. We emphasize macro-F1 as the primary metric because LIAR is imbalanced and macro-F1 averages per-class F1 scores

with equal weight, so it directly reflects performance on minority and borderline classes (e.g., pants-fire, half-true) that would otherwise be dominated by frequent labels under plain accuracy. For each configuration we also report normalized confusion matrices (Sections 4.5 and 4.7) and one-vs-rest ROC and precision–recall curves (Section 4.9) for finer-grained error analysis. All test-set metrics are computed using the checkpoint that achieved the best validation macro-F1.

3.7. Experimental Scope

We deliberately restrict the experimental scope to single-model, text-only transformer fine-tuning, without ensembling, multimodal inputs, or retrieval. All experiments are repeated across the five seeds above and reported as mean ± standard deviation. Paired Wilcoxon signed-rank tests are used to compare Hist ON versus Hist OFF, with the explicit caveat that with $n = 5$ pairs the minimum achievable two-sided p-value is 0.0625 (Section 4.3). This is a known limitation of small-n paired nonparametric testing, not a property of our results.

3.8. Hardware and Software Environment

All experiments were run on CUDA-enabled NVIDIA GPUs. The primary reported runs used an NVIDIA Tesla V100-SXM2 GPU with 32 GB of memory. The software environment used Python 3.9.16, PyTorch 2.7.1 with CUDA 11.8 support, Hugging Face Transformers, pandas, scikit-learn, NumPy, and SciPy. The exact package versions, random seeds, dataset checksums, and run configuration files are recorded in `environment.json` and included with the reproducibility artifacts described in the Data Availability statement.

4. Results

This section reports results for the main LIAR experiments. We use the LIAR test set throughout and only change one factor (Hist OFF versus Hist ON) within each row. All experiments are repeated across five random seeds (7, 13, 42, 123, 2024), and we report mean ± standard deviation; paired Wilcoxon signed-rank tests are reported alongside in Section 4.3.

4.1. Evaluation Setup and Metrics

Models are scored on accuracy and macro-averaged F1 (Macro-F1). Macro-F1 is treated as the primary metric because it gives equal weight to each class, which is more informative under the LIAR label imbalance than plain accuracy. We additionally examine confusion matrices, per-class precision/recall, one-vs-rest ROC curves, and precision–recall curves to understand where errors concentrate, especially around the borderline truthfulness categories.

Notation. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be an evaluation set with N samples and K classes, where $y_i \in \{1, \ldots, K\}$ is the ground-truth label for input x_i and $\hat{y}_i$ is the predicted label. For each class k, we define true positives (TP_k), false positives (FP_k), and false negatives (FN_k) one-vs-rest.

Accuracy.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Precision, Recall, F1 (per class).

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \quad \text{F1}_k = \frac{2\,\text{Precision}_k\,\text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{2}$$

Macro-F1.

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k. \quad (3)$$

Because LIAR is imbalanced and minority categories matter for screening, macro-F1 is our primary metric: it weights every class equally, whereas plain accuracy can be dominated by the most frequent labels. Reporting both lets the reader see whether gains are uniform across classes or concentrated in frequent ones.

Confusion matrices. We report normalized confusion matrices in which each row sums to (approximately) 100% of the true-class mass. Off-diagonal entries indicate systematic confusion between specific categories, which is especially informative for semantically adjacent truthfulness labels (e.g., half-true vs. mostly-true).

One-vs-rest ROC and AUC. For each class k, we compute the ROC curve and its area under the curve AUC_k by treating class k as positive and all others as negative.

Precision–Recall curves and Average Precision. For each class k, we report the one-vs-rest precision–recall curve and its average precision AP_k; under class imbalance, precision–recall curves are usually more informative than ROC curves for minority classes.

Together, accuracy and macro-F1 summarize top-1 performance, confusion matrices localize the errors, and ROC/PR curves describe ranking quality across thresholds.

4.2. Overall Performance Summary

Table 2 summarizes test performance for both label granularities and both backbones as mean ± standard deviation across the five seeds. In every row, Hist ON improves on Hist OFF in both accuracy and macro-F1. The improvement is largest in the six-class setting: for DeBERTa-v3-large, mean test accuracy/macro-F1 increases from $0.3005 \pm 0.0036/0.3096 \pm 0.0062$ (Hist OFF) to $0.4757 \pm 0.0199/0.4773 \pm 0.0235$ (Hist ON); for RoBERTa-large, from $0.3051 \pm 0.0048/0.3127 \pm 0.0071$ to $0.4847 \pm 0.0053/0.4855 \pm 0.0068$. The three-class setting shows clear but smaller absolute gains.

Table 2. LIAR test performance for Hist OFF and Hist ON across task granularities and backbones. Values are mean ± standard deviation over five seeds. Bold indicates the per-task, per-backbone winner.

Task	Hist	Model	Test Acc (Mean ± Std)	Macro-F1 (Mean ± Std)
3-class	OFF	DeBERTa-v3-large	0.5283 ± 0.0237	0.4882 ± 0.0102
3-class	OFF	RoBERTa-large	0.5279 ± 0.0125	0.4907 ± 0.0110
3-class	ON	DeBERTa-v3-large	**0.6618 ± 0.0135**	**0.6241 ± 0.0071**
3-class	ON	RoBERTa-large	0.6156 ± 0.1003	0.5727 ± 0.0944
6-class	OFF	DeBERTa-v3-large	0.3005 ± 0.0036	0.3096 ± 0.0062
6-class	OFF	RoBERTa-large	0.3051 ± 0.0048	0.3127 ± 0.0071
6-class	ON	DeBERTa-v3-large	0.4757 ± 0.0199	0.4773 ± 0.0235
6-class	ON	RoBERTa-large	**0.4847 ± 0.0053**	**0.4855 ± 0.0068**

Quantitatively, enabling speaker history (Hist ON) leads to consistent gains in every configuration: in the 6-class task, DeBERTa improves by +0.1752 in mean accuracy and +0.1677 in mean macro-F1, and RoBERTa by +0.1796 and +0.1728; in the 3-class task, DeBERTa improves by +0.1335 and +0.1359 and RoBERTa by +0.0877 and +0.0820. These gains are visible across all five seeds, with a single exception in the 3-class RoBERTa Hist ON run (seed 13) that we discuss in Section 4.3.

Figure 3 visualizes the controlled comparison. Throughout the figures, we prioritize macro-F1 because it is our primary metric under the class imbalance of LIAR and LIAR2 NEW, as described in Section 5; the corresponding accuracy plots are provided in Appendix A.

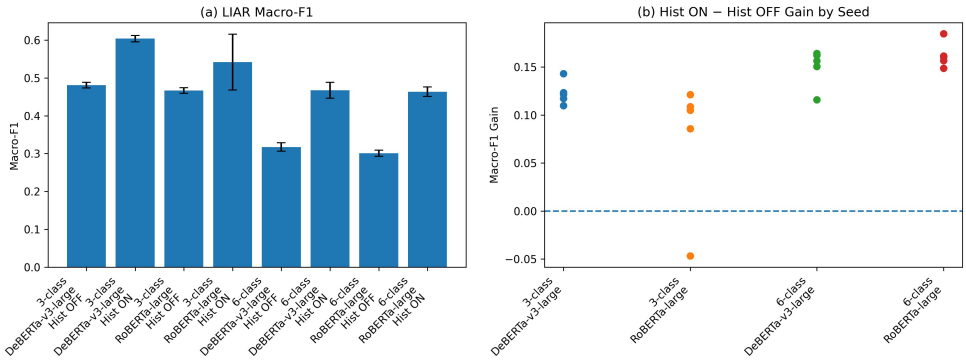


Figure 3. LIAR test performance and seed-level macro-F1 gains. Panel (**a**) reports mean macro-F1 over five seeds, with error bars showing standard deviation. Panel (**b**) shows the seed-level macro-F1 gain from adding speaker-history context, computed as Hist ON minus Hist OFF. Positive values indicate that speaker-history context improves performance. The dashed reference line in Panel (**b**) marks zero gain. The seed-level panel makes the direction of the effect visually clear across backbones and label granularities, while transparently showing the single unstable 3-class RoBERTa-large seed.

4.3. Statistical Reporting: Paired Differences, Effect Sizes, and Seed-Level Gains

To characterize the gains in Table 2 without overclaiming, we report three complementary quantities for the Hist ON versus Hist OFF comparison within each (task, backbone) cell: (i) the paired mean difference over the five seeds, (ii) a paired Wilcoxon signed-rank test (reported with an explicit caveat about its $n = 5$ limit), and (iii) seed-level gains. Table 3 lists the first two.

Table 3. Paired Wilcoxon signed-rank tests comparing Hist ON to Hist OFF within each (task, backbone) cell, over the five seeds. We report the mean paired difference in accuracy and macro-F1 (always with positive sign for Hist ON) and the two-sided p-value. With $n = 5$ paired observations, the minimum achievable two-sided p-value for the Wilcoxon test is 0.0625. The p-values therefore cannot reach the conventional $\alpha = 0.05$ threshold; we report them to make the direction-of-effect explicit, not as evidence of significance in the formal sense.

Task	Model	Acc Diff (Mean)	F1 Diff (Mean)	p (Acc)	p (F1)
6-class	DeBERTa-v3-large	+0.1752	+0.1677	0.0625	0.0625
6-class	RoBERTa-large	+0.1796	+0.1728	0.0625	0.0625
3-class	DeBERTa-v3-large	+0.1335	+0.1359	0.0625	0.0625
3-class	RoBERTa-large	+0.0877	+0.0820	0.1250	0.1250

To make the effect-size analysis visible in the manuscript, Table 4 reports the paired seed-level macro-F1 differences, bootstrap confidence intervals, and Cliff's δ for Hist ON minus Hist OFF.

Table 4. Effect-size and bootstrap reliability summary for Hist ON minus Hist OFF on LIAR. Values are computed from paired seed-level macro-F1 differences across five seeds.

Task	Model	Macro-F1 Diff	95% Bootstrap CI	Cliff's δ
6-class	DeBERTa-v3-large	0.1677	[0.1510, 0.1786]	1.00
6-class	RoBERTa-large	0.1728	[0.1636, 0.1829]	1.00
3-class	DeBERTa-v3-large	0.1359	[0.1246, 0.1472]	1.00
3-class	RoBERTa-large	0.0820	[−0.0119, 0.1385]	0.60

The bootstrap confidence intervals are computed over the paired seed-level macro-F1 differences between Hist ON and Hist OFF. The first three settings show consistently

positive effects, while the 3-class RoBERTa-large setting is more variable because one seed produced a negative Hist ON minus Hist OFF difference.

In seven of the eight (cell, metric) pairs (six of which involve six-class or DeBERTa configurations and the seventh is the 3-class DeBERTa pair), every seed yielded Hist ON $>$ Hist OFF, which produces the minimum achievable p-value of 0.0625. The exception is RoBERTa-large on the 3-class task, where seed 13 produced an unusually low Hist ON test accuracy (0.4400, plausibly because of an unfavorable local minimum during fine-tuning), making the per-seed comparison weaker; the resulting p-value is 0.125 and the mean effect remains positive.

Table 5 shows that the larger standard deviation for the 3-class RoBERTa-large Hist ON setting is mainly driven by seed 13, while the other four seeds remain above 0.60 macro-F1.

Table 5. Per-seed LIAR test results for the 3-class RoBERTa-large Hist ON configuration.

Seed	Test Accuracy	Macro-F1
7	0.6885	0.6332
13	0.4400	0.4053
42	0.6456	0.6165
123	0.6682	0.6012
2024	0.6355	0.6073

With only five paired seeds, the smallest two-sided p-value the Wilcoxon test can produce is 0.0625, so we cannot reach $\alpha = 0.05$ even when all five seeds favor Hist ON. We therefore do not call any of these effects "significant" in the formal sense. What the test does establish is the direction of the effect: in seven of the eight comparisons, every seed favors Hist ON, and the mean differences are positive in all eight comparisons. We emphasize the direction-of-effect and the size of the mean differences as the primary evidence, and report the p-values for transparency.

Effect sizes and seed-level gains. Because p-values are uninformative at $n = 5$, we treat the paired mean differences in Table 3 as the primary effect-size summary. These differences are large relative to the per-seed standard deviations in Table 2; for example, the six-class macro-F1 gain of roughly $+0.17$ is several times larger than the Hist ON standard deviations. To make the reliability analysis visible in the manuscript, Table 4 reports bootstrap confidence intervals and Cliff's δ values for the Hist ON minus Hist OFF macro-F1 differences. These values are interpreted cautiously because they are computed over only five paired seeds. For the transfer evaluation, Figure 4 shows the macro-F1 gain for every individual seed; the gains are positive for the large majority of seeds, and the mean gain is positive in all four configurations.

Prediction-level reliability analysis. Because $n = 5$ seeds limits the Wilcoxon test, we do not claim conventional significance; instead, we supplement seed-level reporting with prediction-level paired analyses computed from the saved test predictions, which use the full test set (thousands of paired examples) rather than five seed-level scores. Specifically, the released analysis code computes on the matched per-example predictions of the Hist ON and Hist OFF checkpoints: (i) a paired bootstrap confidence interval over test examples for the Hist ON minus Hist OFF difference in accuracy and macro-F1; and (ii) a paired prediction test (McNemar's test on the discordant correct/incorrect pairs, with an approximate-randomization test as a distribution-free alternative). These analyses assess the reliability of the observed direction and magnitude of the improvement at the level of individual predictions; they are not a substitute for, and we do not present them as, a claim of universal statistical significance. The per-configuration outputs are written to the released results files so that they are fully reproducible from the saved predictions.

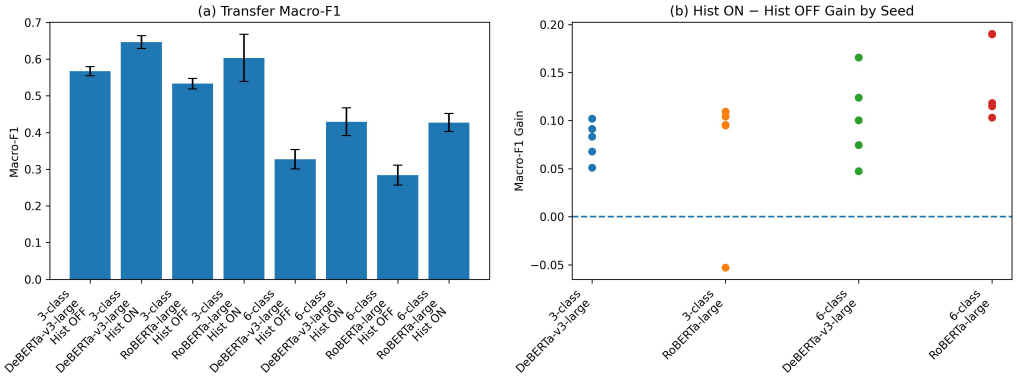
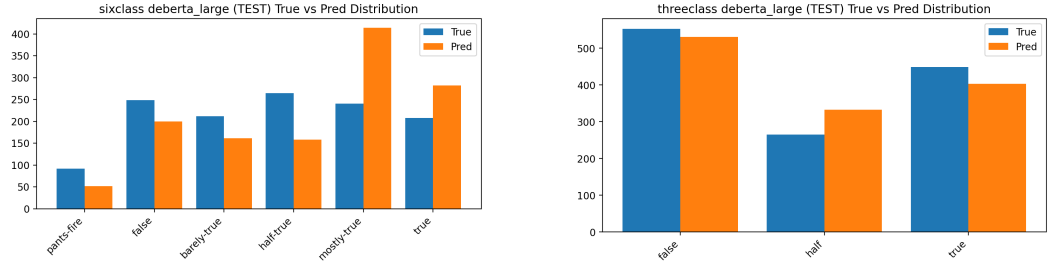


Figure 4. LIAR-to-LIAR2 NEW transfer evaluation. Panel (**a**) reports mean macro-F1 over five LIAR-trained checkpoints, with error bars showing standard deviation across seeds. Panel (**b**) shows the seed-level macro-F1 gain from adding speaker-history context, computed as Hist ON minus Hist OFF for each seed. Positive values indicate that speaker-history context improves transfer performance on the full non-overlapping LIAR2 NEW test split. All four configurations have a positive mean gain; the single negative point (one seed of 3-class RoBERTa-large) mirrors the higher-variance behavior of that configuration already noted on the original LIAR split.

4.4. Class Distribution and Evaluation Context

Figure 5 shows the test-set label distribution under both label spaces. The 6-class setting is visibly imbalanced (pants-fire and true are the smallest categories), which is the main reason we emphasize macro-F1 alongside accuracy: accuracy alone can be inflated by performance on the most frequent labels.



(**a**) Six-class test label distribution (**b**) Three-class test label distribution

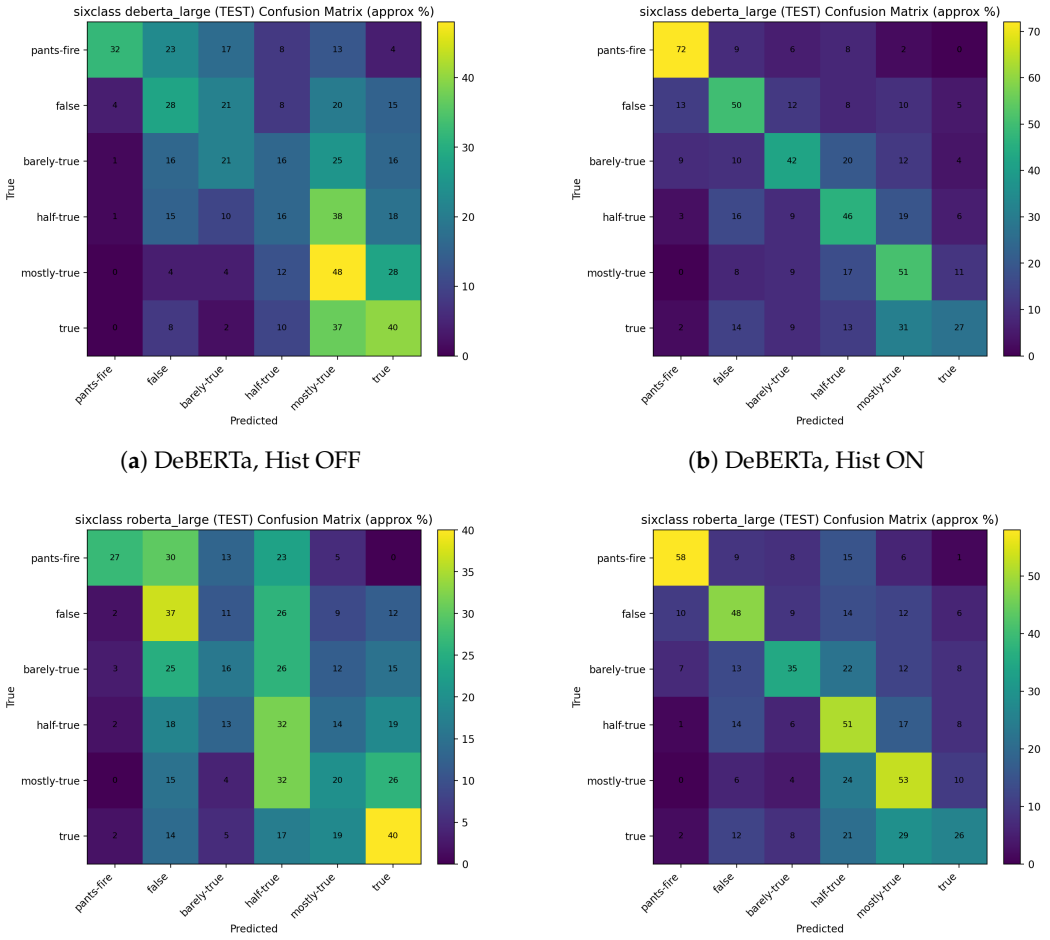
Figure 5. Test-set label distributions for the 6-class (**a**) and 3-class (**b**) LIAR formulations. The imbalance in the fine-grained setting motivates reporting macro-F1 rather than plain accuracy as the primary metric.

4.5. Six-Class Results: Fine-Grained Truthfulness

The 6-class task is the most demanding because several pairs of labels are semantically close (barely-true vs. half-true vs. mostly-true). The best 6-class configuration is DeBERTa-v3-large with Hist ON, with mean test accuracy of 0.4757 ± 0.0199 and macro-F1 of 0.4773 ± 0.0235 over five seeds. RoBERTa-large with Hist ON is comparable (0.4847 ± 0.0053 accuracy, 0.4855 ± 0.0068 macro-F1).

Error analysis (DeBERTa-v3-large, 6-class). Comparing the Hist OFF and Hist ON normalized confusion matrices in Figure 6, the largest reduction in confusion appears around the middle of the truthfulness scale: the diagonal mass for *half-true*, *barely-true*, and *false* grows substantially under Hist ON, and the off-diagonal mass that previously concentrated between *barely-true*/*half-true*/*mostly-true* drops correspondingly. The per-class F1 improvements under Hist ON for DeBERTa are largest for *half-true* (+0.2591), *barely-true* (+0.2005), *false* (+0.1813), and *pants-fire* (+0.1736), with smaller changes for *mostly-true* and

true. This pattern is consistent with the intuition that credibility context is most helpful when the claim text alone is least decisive, i.e., for the middle of the truthfulness scale.



(**a**) DeBERTa, Hist OFF

(**b**) DeBERTa, Hist ON

(**c**) RoBERTa, Hist OFF

(**d**) RoBERTa, Hist ON

Figure 6. Six-class normalized test confusion matrices (approximate %) for Hist OFF (**a**,**c**) versus Hist ON (**b**,**d**), for DeBERTa-v3-large (**a**,**b**) and RoBERTa-large (**c**,**d**). Hist ON consistently increases diagonal mass and reduces systematic confusion between adjacent truthfulness labels, especially around half-true. This matches the per-class F1 improvements reported in the text and explains why macro-F1 improves so much in the fine-grained setting.

Residual error patterns. Even under Hist ON, the residual errors in the 6-class confusion matrices remain concentrated between adjacent truthfulness levels (half-true ↔ mostly-true, mostly-true ↔ true). This is expected: these labels can differ by small factual nuances or framing that are not present in the claim text, and credibility context alone cannot fully resolve them.

Quantitative failure-mode breakdown. To turn the qualitative confusion-matrix discussion into a reproducible analysis, we compute a quantitative failure-mode breakdown directly from the regenerated test predictions of the DeBERTa-v3-large 6-class checkpoints (Table 6 and Figure 7). Each category is defined on the ordered truthfulness scale and computed per seed, then averaged. Hist ON reduces the overall error rate from 69.31% to 54.29%, adjacent-label errors from 36.51% to 28.38% of all examples, over-lenient errors from 41.86% to 29.77%, conservative-bias errors from 26.20% to 22.13%, false-to-half/true transitions from 43.98% to 35.30%, and true-to-half/false transitions from 33.41% to 29.09%. This turns the prior qualitative confusion-matrix discussion into a quantitative, reproducible error analysis. Notably, the adjacent share among the remaining errors is essentially

unchanged (52.69% versus 52.35%): speaker-history context reduces the number of errors, but the errors that remain are still concentrated among neighboring truthfulness labels, which is consistent with the residual-confusion pattern above.

Table 6. Quantitative failure-mode breakdown on the LIAR 6-class test set for DeBERTa-v3-large. Values are reported as mean ± standard deviation over five seeds. Adjacent errors are incorrect predictions within one neighboring truthfulness label. Conservative-bias errors are incorrect predictions that move an extreme label toward the middle of the truthfulness scale. All values are computed from the regenerated test predictions of the saved best-validation checkpoints.

Failure Mode (%)	Hist OFF	Hist ON
Error rate	69.31 ± 1.52	54.29 ± 1.86
Adjacent errors	36.51 ± 0.92	28.38 ± 0.68
Adjacent share of errors	52.69 ± 1.30	52.35 ± 2.61
Over-severe	27.45 ± 1.67	24.51 ± 4.49
Over-lenient	41.86 ± 1.80	29.77 ± 5.73
Conservative bias	26.20 ± 3.54	22.13 ± 3.13
False-to-half/true	43.98 ± 2.22	35.30 ± 10.11
True-to-half/false	33.41 ± 5.47	29.09 ± 9.72

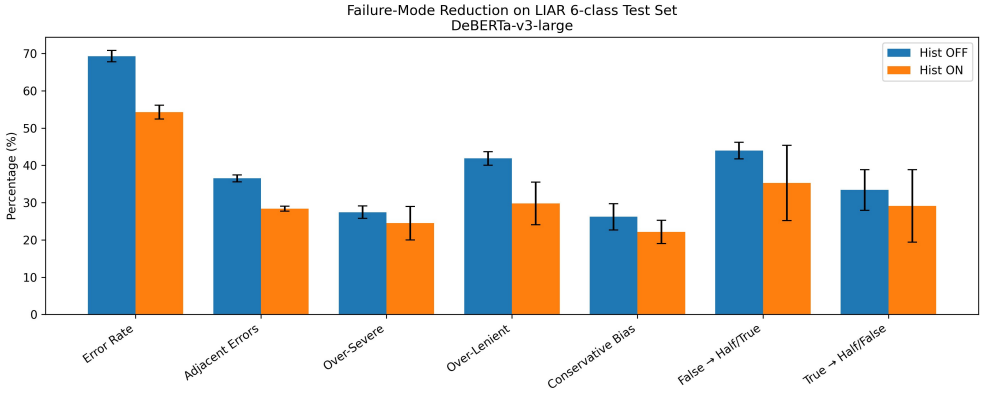


Figure 7. Failure-mode reduction on the LIAR 6-class test set for DeBERTa-v3-large. Bars report mean percentages over five seeds, with error bars showing standard deviation. Lower values indicate fewer errors. Hist ON reduces the overall error rate and several systematic error types, including adjacent-label errors, over-lenient predictions, conservative-bias errors, false-to-half/true transitions, and true-to-half/false transitions. "Adjacent share of errors" from Table 6 is intentionally omitted here to avoid confusion, since it is a share of the (shrinking) error set rather than an absolute rate.

4.6. Comparison with Prior 6-Class LIAR Studies

Within the comparison set used in this paper, we compare against studies that evaluate on the LIAR 6-class label space and whose publication record we could verify. A recent knowledge-graph enhanced model [15] reports 0.4540 accuracy/0.4500 macro-F1 on a 6-class LIAR truthfulness setting.

Our best 6-class configuration (DeBERTa-v3-large, Hist ON) attains mean test accuracy of 0.4757 and macro-F1 of 0.4773 over five seeds (Table 7), which is competitive with this comparison point on the same metric under the same dataset and label space.

We do not claim a state-of-the-art result against the full LIAR literature; we compare only against a study we could verify on the same label mapping: [15] uses a modern transformer-based, knowledge-graph-enhanced pipeline, so it is a meaningful contemporary point of reference. We deliberately do not benchmark against older CNN/attention or basic-BERT LIAR systems as evidence of superiority: such comparisons conflate backbone capacity with the contribution we actually study. The main evidence in this paper is the controlled within-backbone Hist OFF versus Hist ON comparison (Section 4.2)

and the LIAR-to-LIAR2 NEW transfer evaluation (Section 5), both of which hold the architecture fixed.

Table 7. Six-class comparison with prior work, restricted to studies that report on LIAR under the six-class label space. Numbers for prior work are taken from the cited sources; numbers for "Ours" are mean test performance over five seeds with the same pipeline and split. We do not claim state-of-the-art across the full LIAR literature; this table only compares against the specific studies we could verify under the same dataset and label space.

Work	Model/Setting	Signals Used	Test Acc	Macro-F1
[15]	KGFakeNet (6-class setting)	Claim text (+ metadata; no speaker history)	0.4540	0.4500
Ours	DeBERTa-v3-large (Hist OFF) (LIAR, 6-class)	Claim text (no speaker history)	0.3005	0.3096
Ours	DeBERTa-v3-large (Hist ON) (LIAR, 6-class)	Claim text + speaker history counts	0.4757	0.4773

4.7. Three-Class Results: Coarse-Grained Truthfulness

The 3-class mapping ({false, half, true}) is a common simplification that reduces ambiguity while preserving meaningful distinctions between misleading, partially correct, and accurate statements. Under this mapping, Hist ON still provides clear gains: the best 3-class configuration is DeBERTa-v3-large with Hist ON, with mean test accuracy of 0.6618 ± 0.0135 and macro-F1 of 0.6241 ± 0.0071 across five seeds. Compared to DeBERTa with Hist OFF, mean macro-F1 increases from 0.4882 to 0.6241.

Across both backbones, the per-class improvement is concentrated on the "half" class, which sits between the other two and is the hardest to judge from claim text alone: for DeBERTa, the "half" F1 increases by $+0.1560$ under Hist ON, and for RoBERTa by $+0.1260$. This is consistent with the six-class pattern: credibility context is most helpful at the borderline categories where the claim text alone is ambiguous.

4.8. Three-Class Results in Context

We report our 3-class results (Table 2) without a prior-work comparison table, because we were unable to verify the publication record of the older LIAR 3-class baselines we initially considered, and we do not cite unverifiable sources. As in the 6-class case, the primary evidence for the value of credibility context is internal and controlled: the within-backbone Hist OFF versus Hist ON comparison (Section 4.2) and the LIAR-to-LIAR2 NEW transfer evaluation (Section 5), both of which hold the architecture fixed. Our best 3-class configuration (DeBERTa-v3-large, Hist ON) attains mean test accuracy of 0.6618 and macro-F1 of 0.6241 over five seeds.

4.9. Ranking Quality (ROC and Precision–Recall)

ROC and precision–recall curves describe ranking quality across decision thresholds, rather than only the top-1 prediction, which is useful under label imbalance. To keep the imbalance-aware evaluation visible in the main text, Table 8 reports macro-averaged ROC-AUC and average precision for the DeBERTa-v3-large Hist ON checkpoints used for the main ROC and precision–recall analyses.

The one-vs-rest ROC and precision–recall curves, together with the accuracy figures, are provided in Appendix A. They show the same qualitative pattern as the macro-F1 results: Hist ON improves ranking quality, particularly for the borderline categories.

Table 8. Macro-averaged ROC–AUC and average precision for DeBERTa-v3-large under Hist ON on LIAR. Values are computed from the seed 42 test predictions.

Task	Macro ROC–AUC	Macro-Average Precision
6-class	0.8176	0.5591
3-class	0.8093	0.6917

5. LIAR-to-LIAR2 NEW Transfer Evaluation

The results in Section 4 establish that speaker-history credibility context helps on the original LIAR test split. A natural follow-up question is whether this benefit is specific to that particular split or whether it survives on a related, non-overlapping benchmark. To answer this, we run a transfer evaluation: we take the LIAR-trained checkpoints from Section 4 and evaluate them directly on the LIAR2 NEW test set, with no additional fine-tuning on LIAR2. This is transfer validation of the existing checkpoints, not retraining on LIAR2.

We use LIAR2 NEW as the related transfer benchmark because it preserves the key structure needed for this study: political truthfulness labels, speaker identity, and speaker-history count fields [20]. These fields allow the same Hist OFF versus Hist ON comparison used on LIAR to be evaluated on a non-overlapping related benchmark. This design lets us test whether the credibility-context benefit transfers beyond the original LIAR test split while keeping the evaluation aligned with the speaker-history hypothesis of the study.

5.1. Dataset, Filtering, and Leakage Prevention

LIAR2 [20] is a revised and expanded political claim dataset that re-uses the LIAR truthfulness scale and retains speaker metadata, including speaker fact-checking history counts. Because LIAR2 partially overlaps with LIAR at the statement level, we construct a NEW portion: we remove from the LIAR2 test split any statement that also appears in LIAR (across LIAR train, validation, and test). After this filtering, the LIAR2 NEW test split has zero statement-level overlap with LIAR, and contains 1098 test examples. This filtering is what makes the transfer evaluation informative: a model trained on LIAR has never seen any of these statements during training. The filtering rule, and an assertion that aborts if any overlap remains, are included in the released code.

Table 9 summarizes the two datasets and their roles in this study.

Table 9. Datasets used in this study. LIAR is the training and primary-evaluation benchmark; LIAR2 NEW is used only for transfer evaluation of the LIAR-trained checkpoints without fine-tuning on LIAR2 NEW.

Dataset	Number of Examples	Number of Labels	Context/Metadata Fields	Speaker History	Role
LIAR [9]	12,836 (train/valid/test)	6 (mapped to 3)	subject, speaker, job, state, party, context	Yes (5 numeric counts)	Training + primary evaluation
LIAR2 NEW [20]	1098 (test only; non-overlapping)	6 (mapped to 3 by the same rule)	subject, speaker, job, state, party, context	Yes (numeric counts)	Transfer evaluation only

5.2. Evaluation Design

The transfer evaluation reuses, without modification, the five LIAR-trained checkpoints (one per seed) for each (task, backbone, Hist setting) cell from Section 4. For each checkpoint, we build the LIAR2 NEW test inputs with exactly the same Hist OFF/Hist

ON construction described in Section 3.2 (Hist ON appends the LIAR2 speaker-history counts as the structured `speaker_history` field; Hist OFF omits them), tokenize identically, and compute accuracy and macro-F1 on the 1098-example LIAR2 NEW test split. We report mean ± standard deviation over the five seeds, exactly as in the main experiments. The 6-to-3 label mapping is unchanged. No parameter of any checkpoint is updated during this evaluation.

5.3. Transfer Results

Table 10 reports the transfer results, and Figure 4 visualizes them. In every one of the four backbone/granularity settings, Hist ON improves macro-F1 over Hist OFF on the non-overlapping LIAR2 NEW test split. The paired macro-F1 transfer gains are +0.1024 (6-class DeBERTa-v3-large), +0.1433 (6-class RoBERTa-large), +0.0791 (3-class DeBERTa-v3-large), and +0.0704 (3-class RoBERTa-large). The corresponding accuracy gains are +0.0940, +0.1306, +0.0710, and +0.0674. The strongest transfer gain is for six-class RoBERTa-large, where macro-F1 increases from 0.2839 to 0.4272.

Table 10. LIAR-to-LIAR2 NEW transfer performance using LIAR-trained checkpoints without additional fine-tuning. Values are mean ± standard deviation over five seeds. Bold indicates the per-task, per-backbone winner.

Task	Hist	Model	Test Acc (Mean ± Std)	Macro-F1 (Mean ± Std)
6-class	OFF	DeBERTa-v3-large	0.3599 ± 0.0374	0.3269 ± 0.0264
6-class	OFF	RoBERTa-large	0.3155 ± 0.0277	0.2839 ± 0.0272
6-class	ON	DeBERTa-v3-large	**0.4539 ± 0.0430**	**0.4293 ± 0.0378**
6-class	ON	RoBERTa-large	**0.4461 ± 0.0292**	**0.4272 ± 0.0244**
3-class	OFF	DeBERTa-v3-large	0.7408 ± 0.0267	0.5671 ± 0.0123
3-class	OFF	RoBERTa-large	0.7157 ± 0.0215	0.5329 ± 0.0146
3-class	ON	DeBERTa-v3-large	**0.8118 ± 0.0207**	**0.6462 ± 0.0173**
3-class	ON	RoBERTa-large	**0.7831 ± 0.0667**	**0.6033 ± 0.0639**

The higher 3-class Hist OFF accuracy on LIAR2 NEW compared with the original LIAR test split should not be interpreted as direct evidence that LIAR2 NEW is easier in all respects. The two test sets differ in label distribution, filtering, annotation process, and the coarser three-class mapping. In particular, accuracy is sensitive to class balance, whereas macro-F1 remains the primary metric for comparing behavior under imbalance. We therefore interpret the LIAR2 NEW results mainly through paired Hist ON versus Hist OFF differences within the same LIAR2 NEW test split, rather than through raw accuracy comparisons across LIAR and LIAR2 NEW.

Table 11 reports the speaker-level overlap between the official LIAR training split and the LIAR2 NEW test split.

Table 11. Speaker-level overlap between the official LIAR training split and the LIAR2 NEW test split.

Quantity	Value
Unique speakers in LIAR train	2910
Unique speakers in LIAR2 NEW test	418
Overlapping speakers	172
LIAR2 NEW test examples with LIAR-train speaker	747
Share of LIAR2 NEW test examples with LIAR-train speaker	68.03%

This speaker-level overlap shows that the LIAR2 NEW transfer split is non-overlapping at the statement level but not speaker-disjoint. Therefore, the LIAR2 NEW results

should be interpreted as related-benchmark transfer rather than as a fully speaker-independent evaluation. This distinction is important because speaker-history features may still benefit from recurring speakers across datasets. For this reason, we interpret the LIAR2 NEW results primarily through paired Hist ON versus Hist OFF differences within the same LIAR2 NEW test split, rather than as evidence of fully speaker-independent generalization.

The seed-level gain plot in Figure 4b shows that the macro-F1 gain is positive for the large majority of individual seeds, not just on average. The only negative seed-level gain occurs in one seed of the 3-class RoBERTa-large configuration, which is the same configuration that showed the highest cross-seed variance on the original LIAR split (Section 4.2); the configuration's mean transfer gain remains clearly positive (+0.0704).

5.4. Interpretation

The LIAR-to-LIAR2 NEW transfer evaluation provides related-benchmark evidence that the benefit of speaker-history credibility context is not limited to the original LIAR test split. Across both backbones and both label granularities, Hist ON improves macro-F1 over Hist OFF. The strongest transfer gain appears in the 6-class RoBERTa setting, where macro-F1 increases from 0.2839 to 0.4272. Because LIAR2 NEW is related to LIAR in task design and metadata structure, these results should be interpreted as related-benchmark transfer rather than universal out-of-domain generalization. In particular, LIAR2 NEW shares the truthfulness-label scale and the speaker-history count structure of LIAR, so a model that learned to use speaker history on LIAR can apply the same mechanism here; the transfer result says that mechanism keeps working on unseen statements, not that it would work on an arbitrarily different fact-checking domain.

6. Discussion

This study isolates the contribution of speaker history counts as a credibility-context signal, holding the backbone model, data split, and training pipeline constant, and reporting all results over five random seeds. The discussion below is organized around the headline finding, the conceptual distinction between classification and verification, where the gains are concentrated, and the risks that the design surfaces.

6.1. Main Finding

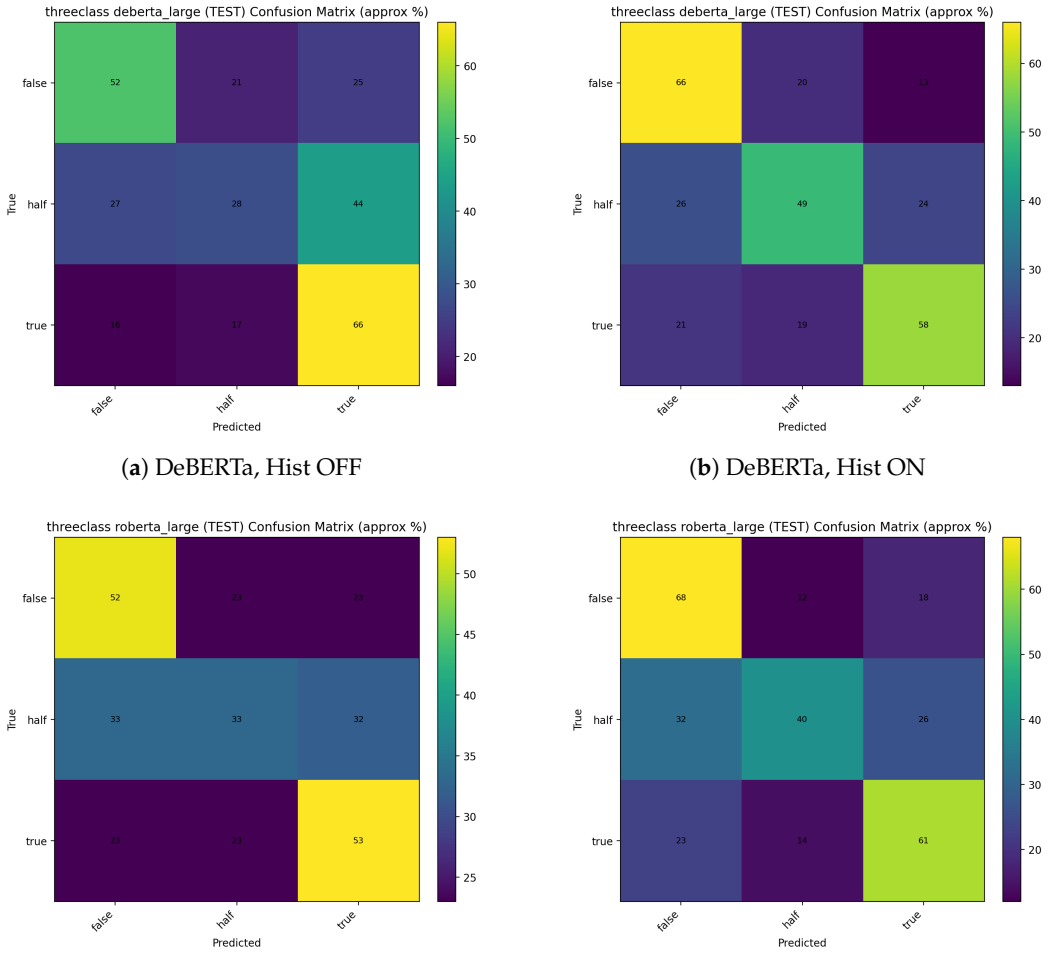
Across both backbones, enabling speaker history (Hist ON) produces consistent gains in both accuracy and macro-F1 in every (task, backbone) cell. In the 6-class setting, mean macro-F1 improves by approximately +0.168 to +0.173 and mean accuracy improves by approximately +0.175 to +0.180. In the 3-class setting, mean macro-F1 improves by approximately +0.082 to +0.136 and mean accuracy improves by approximately +0.088 to +0.134. The gains are larger in the 6-class task, as expected: fine-grained truthfulness labels are more ambiguous and harder to separate from claim text alone, so an additional disambiguation signal helps more there.

6.2. Classification, Not Evidence-Based Verification

We frame our contribution as credibility-augmented classification, not as evidence-grounded verification. The system does not retrieve external evidence and cannot produce a justified verdict that points to specific supporting documents. Hist ON augments the input with a structured behavioral prior over the speaker, which is conceptually similar to how a human fact-checker might raise their scrutiny threshold when a known repeat offender makes a new claim. That kind of prior is useful for prioritization and screening (deciding which claims warrant deeper investigation), but it is not a substitute for evidence-based fact-checking when an evidence-backed verdict is required.

6.3. Why the Gains Concentrate in Borderline Categories

A robust pattern across both backbones is that the largest per-class improvements under Hist ON appear in the middle of the truthfulness scale: half-true, barely-true, and false in the 6-class task, and the half class in the 3-class task. These are precisely the categories where the claim text alone is least decisive: a short statement may sound plausible without being fully accurate, and the labels themselves differ from each other by degree rather than by kind. Speaker credibility history acts as a prior that pushes the model toward the appropriate side of the ambiguity, which is consistent with the confusion-matrix changes in Figures 6 and 8.



(**a**) DeBERTa, Hist OFF (**b**) DeBERTa, Hist ON

(**c**) RoBERTa, Hist OFF (**d**) RoBERTa, Hist ON

Figure 8. Three-class normalized test confusion matrices (approximate %) for Hist OFF (**a**,**c**) versus Hist ON (**b**,**d**), for DeBERTa-v3-large (**a**,**b**) and RoBERTa-large (**c**,**d**). The biggest visible change under Hist ON is that the diagonal mass on the half row increases and the spillover into the neighboring false and true columns is reduced, which is the same pattern as in the 6-class case at a coarser granularity.

6.4. Error Patterns and Qualitative Behavior

Even under Hist ON, the residual errors in the 6-class confusion matrices remain concentrated between adjacent truthfulness levels (e.g., true vs. mostly-true, half-true vs. mostly-true). This is consistent with the intuition that these labels differ by small factual or framing details that are not in the claim text, and so cannot be resolved by credibility context alone.

The same mechanism that helps in borderline cases can mislead in some cases. Two qualitative behaviors are worth flagging. These examples are drawn from the seed-42

DeBERTa-v3-large 6-class Hist ON run for readability and should be interpreted as illustrative rather than exhaustive. The broader patterns they illustrate are supported by the multi-seed confusion matrices and failure-mode summaries reported in Section 4.5:

- Helpful case. A claim labeled half-true from a speaker with a high count of prior half-true and barely-true ratings is sometimes predicted as mostly-true under Hist OFF because the claim text is plausible-looking, but is correctly classified as half-true under Hist ON. The credibility prior nudges the model toward a more cautious assessment.
- Misleading case. A factually true claim from a speaker with predominantly negative credibility history (many prior false or pants-fire ratings) is sometimes shifted from a correct true prediction under Hist OFF to a more conservative mostly-true under Hist ON. The negative credibility prior in this case introduces a conservative bias that overrides the textual evidence and produces an error that the Hist OFF model did not make.

A side-effect of this conservative bias is that Hist ON sometimes makes the true region of the prediction space more cautious. In a downstream triage or screening setting this is often acceptable (it errs toward additional human review for claims by historically unreliable speakers), but in a system intended to produce final verdicts, it would be problematic and should be flagged in deployment.

6.5. Shortcut Learning Risk

A standard concern with appending metadata as plain text is that the model may learn to rely on the history field as a shortcut, effectively underweighting the claim text when history is informative. The Hist ON vs Hist OFF comparison demonstrates that history features carry genuine signal, but it does not by itself rule out shortcut learning. Stronger evidence would come from (i) attribution analyses (e.g., gradient-based attribution or attention rollout) that quantify how much each input field drives the prediction; (ii) counterfactual experiments in which the same claim is re-attributed to speakers with different history profiles, to measure how much the model's prediction depends on the speaker identity versus the claim itself; and (iii) speaker-disjoint evaluation splits (cold-start scenarios) to test whether the model degrades gracefully when the speaker history is uninformative or absent.

6.6. Practical Implication

Speaker history counts are lightweight at inference because they add only five structured count values per claim. The Hist ON variant adds only a small number of structured tokens to the input and does not change the transformer architecture, so its computational overhead is limited to slightly longer tokenized sequences. We did not conduct a dedicated runtime benchmark, so we treat computational efficiency as a practical design motivation rather than as a measured contribution. This makes credibility context attractive for large-scale screening pipelines where retrieval-based verification may not run within the latency or cost budget. We stress that this argument is about screening (deciding what to look at next), not about producing a final verdict for a downstream user.

7. Limitations

We list the limitations we believe are most relevant for a reader assessing the strength of the conclusions.

Primary evaluation on related benchmarks. Training is conducted on a single benchmark (LIAR), and the transfer evaluation in Section 5 uses LIAR2 NEW, which is deliberately related to LIAR (it shares the truthfulness-label scale and the speaker-history count structure). The transfer result is therefore evidence of robustness on a related, non-overlapping

benchmark, not of universal out-of-domain generalization. Evaluation on a benchmark with a genuinely different domain or labeling convention remains future work.

Speaker history can act as a shortcut. The Hist ON setting gives the model access to a strong speaker-level prior. This is useful in distribution but can fail on (i) speakers with little or no history, (ii) shifts in the political environment that change a speaker's typical truthfulness pattern, and (iii) speakers whose history reflects past coverage decisions by fact-checkers rather than an inherent property of the speaker. The conservative-bias error described above is an instance of this risk.

No external evidence retrieval. The approach does not retrieve and cannot produce evidence-backed explanations. Where a downstream user needs a justified verdict with citations, this method is not a substitute for an evidence-grounded fact-checker.

Bias, fairness, and subgroup reliability. Because LIAR does not provide demographic attributes, we do not perform demographic fairness analysis and make no demographic fairness claims. Given the available metadata, the appropriate analysis is subgroup reliability: whether the credibility-context benefit is uniform across metadata-defined groups such as political party, speaker-history density (high- versus low-history speakers), and speaker frequency (frequent versus sparse speakers in training). Speaker history itself reflects past coverage and labeling decisions, which can encode annotator and editorial biases, so a model that relies on this signal could behave unevenly across these groups. We did not compute this subgroup-reliability breakdown for the present revision, so we report it honestly as a limitation rather than as a completed result. We therefore leave party-level, speaker-frequency, and history-density subgroup reliability analysis as future work.

Limited generalization to unseen speakers. The LIAR split is not speaker-disjoint, so a model that uses speaker identity (directly or indirectly through history counts) can benefit from speaker overlap between train and test. The claim about Hist ON benefit therefore needs to be retested on speaker-disjoint splits before it can be trusted to generalize to genuinely new public figures.

Small number of random seeds. Five seeds are enough to detect a consistent directional effect but are not enough to achieve $\alpha = 0.05$ significance under a paired Wilcoxon test (the minimum two-sided p-value with $n = 5$ is 0.0625). We report this explicitly in Section 4.3 and we do not call any of the effects "significant" in the formal sense.

Sensitivity in one configuration. The 3-class RoBERTa-large Hist ON setting shows higher cross-seed variance than the others ($\sigma \approx 0.10$ for accuracy), driven by one seed that converged to a poor solution. We have not yet diagnosed whether this is specific to the optimizer landscape under that configuration or a more general fragility; it is worth follow-up.

8. Future Work

The natural next steps build directly on the limitations listed above.

- Ensemble learning for stronger overall performance. A direct extension is to combine complementary models, such as DeBERTa-v3-large and RoBERTa-large, through ensemble learning. This could include probability averaging, weighted voting, stacking, or seed-level checkpoint ensembling. Such an approach may improve overall accuracy, macro-F1, and stability across random seeds while preserving the same Hist ON versus Hist OFF comparison framework.
- Justification-aware ablation on LIAR-PLUS. An optional follow-up is to use the LIAR-PLUS justifications as a complementary signal alongside, or in place of, the speaker-history counts, to measure how much of the Hist ON gain is recoverable from human-written justifications.

- Speaker-disjoint splits. Reconstruct the LIAR splits so that no speaker appears in both train and test, and re-run the Hist ON versus Hist OFF comparison. A drop in Hist ON benefit on speaker-disjoint splits would provide evidence of speaker-shortcut learning.
- Counterfactual speaker-history experiments. Re-attribute fixed claims to speakers with different history profiles while holding the claim text constant, to quantify how much the prediction depends on the history field.
- Attribution and attention analysis. Use gradient-based attribution or attention rollout to measure which input fields drive the prediction under Hist ON, especially for borderline classes.
- Better metadata fusion. Replace plain-text concatenation of history counts with a structured metadata encoder or attention-based fusion module, and test whether the Hist ON gain increases, decreases, or becomes more interpretable. A structured encoder may provide clearer separation between claim text and metadata than text concatenation alone.

9. Conclusions

We presented a controlled study of credibility context for political claim classification on LIAR. Holding the backbone, data split, and training pipeline fixed, and varying only whether the five LIAR speaker history counts are appended to the input as structured text, we observed consistent gains in both accuracy and macro-F1 across both DeBERTa-v3-large and RoBERTa-large and across both 6-class and 3-class formulations. The gains were largest in the 6-class setting (mean macro-F1 increased by $+0.168$ to $+0.173$) and were concentrated, at the per-class level, in the borderline truthfulness categories where claim text alone is least decisive. Confusion-matrix analyses showed reduced systematic confusion between adjacent labels, and ROC and precision–recall analyses showed improved ranking quality under class imbalance.

We were careful to frame the result as credibility-augmented truthfulness classification, not as evidence-grounded fact verification: the system does not retrieve external evidence and cannot produce a justified verdict. Its natural deployment role is as a low-cost screening or triage signal, not as a replacement for evidence-backed fact-checking. We also showed, through a LIAR-to-LIAR2 NEW transfer evaluation, that the speaker-history benefit is not specific to the original LIAR test split: transferring the LIAR-trained checkpoints to the non-overlapping LIAR2 NEW test set improved macro-F1 under Hist ON in all four backbone/granularity settings. Because LIAR2 NEW is related to LIAR, we read this as related-benchmark robustness rather than universal generalization. Finally, we listed the limitations we consider most important (training on a single benchmark, related-benchmark rather than out-of-domain transfer, speaker-shortcut risk, no retrieval, bias and fairness, limited generalization to unseen speakers, and small-n statistical power) so that the reader can calibrate the strength of the conclusions accordingly.

Author Contributions: Conceptualization, B.A.; methodology, B.A.; software, B.A.; validation, B.A. and F.S.; formal analysis, B.A. and F.S.; investigation, B.A. and F.S.; resources, B.A.; writing—original draft preparation, B.A.; writing—review and editing, B.A. and F.S.; visualization, B.A. and F.S.; supervision, F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The LIAR dataset is publicly available from the original release of [9]: https://www.cs.ucsb.edu/~william/data/liar_dataset.zip (accessed on 25 December 2025). The LIAR2 dataset is publicly available from [20] through Hugging Face: https://huggingface.co/datasets/chengxuphd/liar2 (accessed on 5 April 2026). No new datasets were created in this study. The LIAR2 NEW transfer split was derived from LIAR2 by removing statements that also appear

in the original LIAR train, validation, or test splits. The code used for preprocessing, split construction, training, evaluation, statistical analysis, and figure generation will be made available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary Figures

We collect here the secondary figures referenced in the main text. Macro-F1 is our primary metric because LIAR and LIAR2 NEW are class-imbalanced; accuracy and threshold-based ranking curves are reported for completeness.

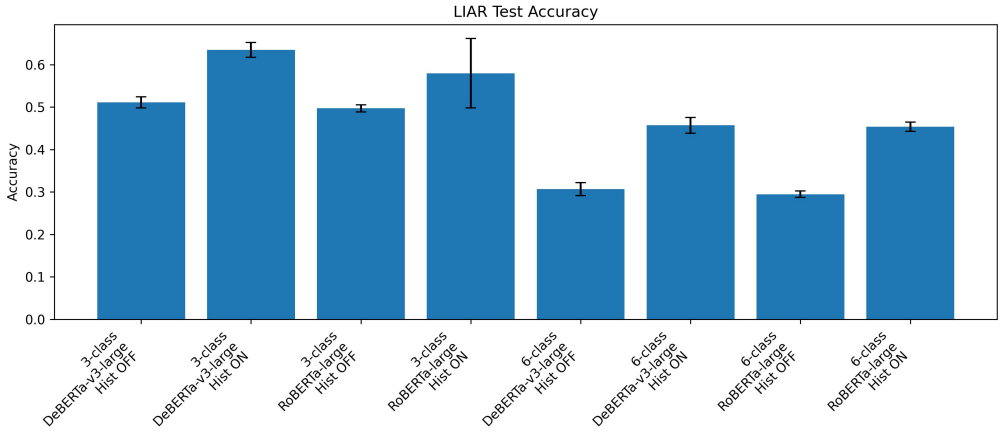


Figure A1. Main LIAR test accuracy (mean over five seeds, error bars showing standard deviation), corresponding to the macro-F1 results in Figure 3. Accuracy is reported for completeness; macro-F1 is the primary metric under class imbalance.

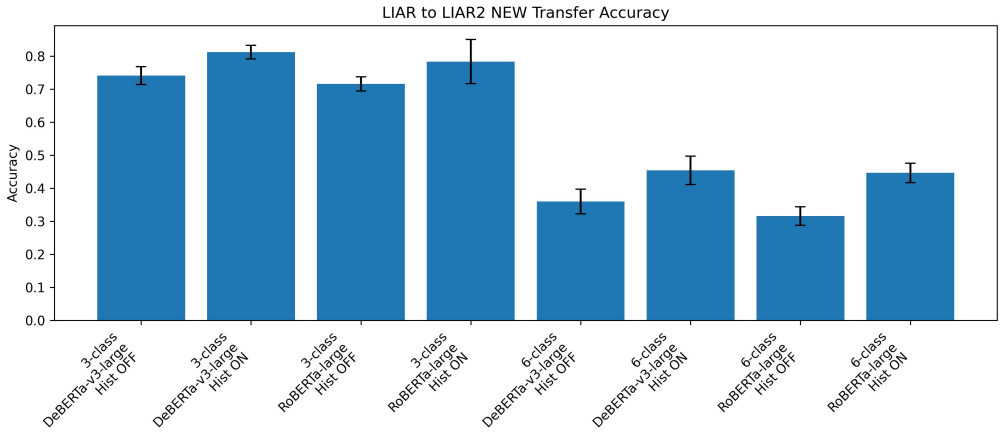
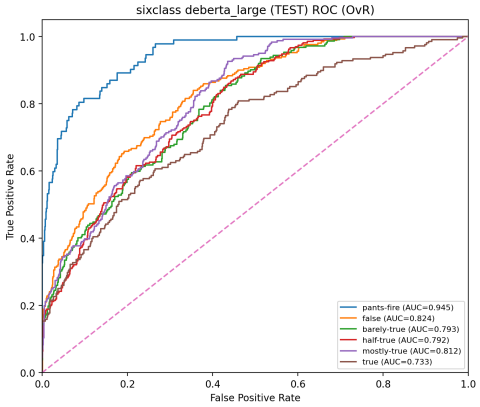
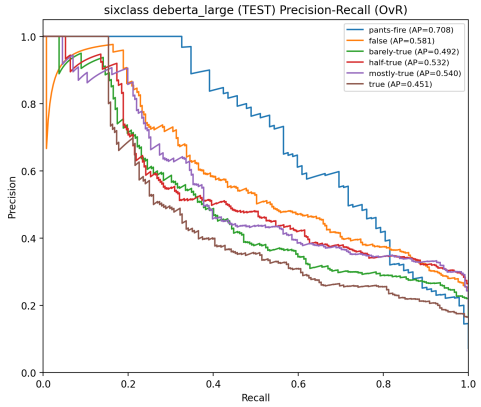


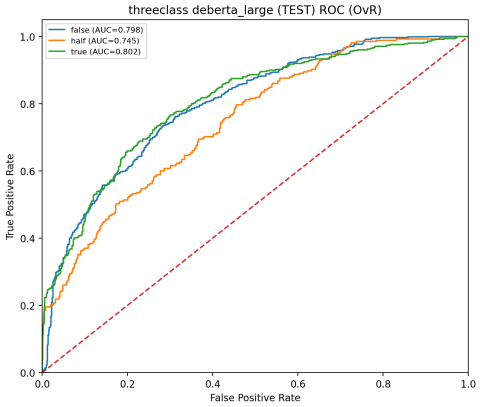
Figure A2. LIAR-to-LIAR2 NEW transfer accuracy (mean over five LIAR-trained checkpoints, error bars showing standard deviation), corresponding to the transfer macro-F1 results in Figure 4. Accuracy is reported for completeness; macro-F1 is the primary metric under class imbalance.



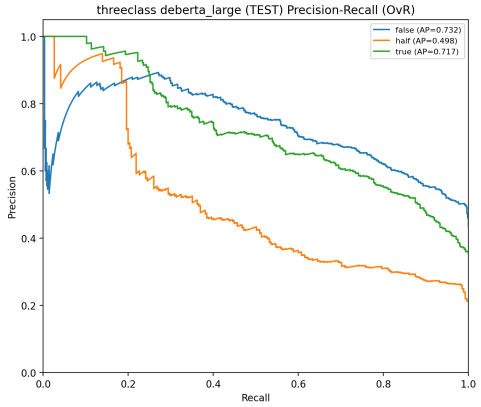
(**a**) Six-class OvR ROC (DeBERTa, Hist ON)



(**b**) Six-class OvR PR (DeBERTa, Hist ON)



(**c**) Three-class OvR ROC (DeBERTa, Hist ON)



(**d**) Three-class OvR PR (DeBERTa, Hist ON)

Figure A3. One-vs-rest ROC (**a**,**c**) and precision–recall (**b**,**d**) curves on the LIAR test set for DeBERTa-v3-large under Hist ON, in the 6-class (**a**,**b**) and 3-class (**c**,**d**) settings. These curves complement accuracy and macro-F1 by showing ranking quality across thresholds, which is especially relevant under class imbalance and for borderline truthfulness categories.

References

1. Okechukwu, C. Media Influence on Public Opinion and Political Decision-Making. *Int. J. Political Sci. Stud.* **2023**, *1*, 13–24.
2. Alkomah, B.; Sheldon, F. Advancements in fake news detection using machine and deep learning models: Comprehensive literature review. In *Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI)*; IEEE: Piscataway, NJ, USA, 2023; pp. 845–852.
3. Micallef, N.; Armacost, V.; Memon, N.; Patil, S. True or false: Studying the work practices of professional fact-checkers. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–44. [CrossRef]
4. Juneja, P.; Mitra, T. Human and technological infrastructures of fact-checking. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–36. [CrossRef]
5. Nakov, P.; Corney, D.; Hasanain, M.; Alam, F.; Elsayed, T.; Barrón-Cedeño, A.; Papotti, P.; Shaar, S.; Da San Martino, G. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence Organization: Palo Alto, CA, USA, 2021; pp. 4551–4558. [CrossRef] [PubMed]
6. Guo, Z.; Schlichtkrull, M.; Vlachos, A. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 178–206. [CrossRef]
7. Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G.L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **2024**, *6*, 852–863. [CrossRef]
8. Konstantinovskiy, L.; Price, O.; Babakar, M.; Zubiaga, A. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digit. Threat. Res. Pract.* **2021**, *2*, 1–16. [CrossRef]

9. Wang, W.Y. "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 (Volume 2: Short Papers), Vancouver, BC, Canada*; Barzilay, R., Kan, M.Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 422–426. [CrossRef]
10. Fetzer, A. Context: Theoretical analysis and its implications for political discourse analysis. In *Handbook of Political Discourse*; Edward Elgar Publishing: Cheltenham, UK, 2023; pp. 164–179.
11. Moravec, P.L.; Kim, A.; Dennis, A.R.; Minas, R.K. Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *Inf. Syst. Res.* **2022**, *33*, 887–907. [CrossRef]
12. Srba, I.; Razuvayevskaya, O.; Leite, J.A.; Moro, R.; Schlicht, I.B.; Tonelli, S.; García, F.M.; Lottmann, S.B.; Teyssou, D.; Porcellini, V.; et al. A survey on automatic credibility assessment using textual credibility signals in the era of large language models. *ACM Trans. Intell. Syst. Technol.* **2026**, *17*, 1–80. [CrossRef]
13. Alghamdi, J.; Lin, Y.; Luo, S. Machine learning and deep learning approaches for fake news detection and related topics in multilingual contexts: A systematic literature review. *Multimed. Tools Appl.* **2026**, *85*, 353. [CrossRef]
14. Panda, S.; Levitan, S.I. Deception Detection Within and Across Domains: Identifying and Understanding the Performance Gap. *ACM J. Data Inf. Qual.* **2022**, *15*, 1–27. [CrossRef]
15. Kumar, A.; Kumar, P.; Yadav, A.; Ahlawat, S.; Prasad, Y. KGFakeNet: A Knowledge Graph-Enhanced Model for Fake News Detection. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK), Abu Dhabi, United Arab Emirates*; Gesese, G.A., Sack, H., Paulheim, H., Merono-Penuela, A., Chen, L., Eds.; International Committee on Computational Linguistics: New York, NY, USA, 2025; pp. 109–122.
16. Arora, S.; Agrawal, V.; Kumar, D.; Arora, S.; Banshal, S.K. Sentimental impact of fake news on social media using an integrated ensemble framework. *Soc. Netw. Anal. Min.* **2024**, *14*, 185. [CrossRef]
17. Liu, H.; Wang, W.; Li, H.; Li, H. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Proceedings of the Findings of the Association for Computational Linguistics*; ACL: Stroudsburg, PA, USA, 2024; pp. 15556–15583.
18. Glockner, M.; Staliūnaitė, I.; Thorne, J.; Vallejo, G.; Vlachos, A.; Gurevych, I. AmbiFC: Fact-Checking Ambiguous Claims with Evidence. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 1–18. [CrossRef]
19. Calvo Figueras, B.; Cuadros Oller, M.; Agerri, R. A Semantics-Aware Approach to Automated Claim Verification. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022.
20. Xu, C.; Kechadi, T. An Enhanced Fake News Detection System with Fuzzy Deep Learning. *IEEE Access* **2024**, *12*, 88006–88021. [CrossRef]
21. Biçer, B.; Durmaz, B.; Aras, S.; Dibeklioğlu, H. DECEPTIcON: Bridging Gaps in In-the-Wild Deception Research. *IEEE Trans. Affect. Comput.* **2025**, *16*, 3452–3464. [CrossRef]