

CAWE-X: A Confidence-Aware Explainable Weighted Ensemble Framework for Misleading Online Content Detection with Calibrated Uncertainty

Muntaha abdulzahra hatem

Ministry of Higher Education and Scientific Research – Department of Scholarships and Cultural Relations.
Email: muntaha.hatem@moheer.edu.iq

ABSTRACT

Detection of potentially deceptive content from the web is generally considered as the choice of one optimal classifier, but in practice, several classifiers based on the same representation produce nearly equal performance while having different classifications on particular objects, have poor prediction confidence and cannot provide explanations for their decisions. This work introduces the CAWE-X model which is the Confidence-Aware Explainable Weighted Ensemble approach which combines four complementary models – Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Machine and Random Forest using TF-IDF n-gram representation – into one collaborative classification mechanism. Calibration of the scores of the learners is performed using either Platt scaling or isotonic regression based on the out-of-fold predictions, while weighting instances is a combination of validation skills of the learners and their prediction confidence. The framework outputs a Final Misleading Score, abstains with a Requires Verification label when the fused confidence is low or the learners are evenly split, and explains every decision through ensemble-weighted n-gram attributions and per-learner contribution shares. The framework is evaluated on LIAR and FakeNewsNet (PolitiFact and GossipCop) with leakage control, staged ablation, cross-dataset validation, a DistilBERT baseline, bootstrap confidence intervals, statistical tests, calibration analysis, risk-coverage analysis, explanation faithfulness testing and structured error analysis. In the illustrative LIAR results, CAWE-X achieved a macro-F1 of 0.683 and ROC-AUC of 0.756, compared with 0.660/0.729 for the strongest single conventional learner (SVM) and 0.692/0.771 for DistilBERT. Calibration reduced ensemble ECE from 0.049 to 0.026. At the selected abstention threshold ($\tau = 0.10$), coverage was 0.801 and selective macro-F1 increased to 0.733. Cross-dataset evaluation showed the expected degradation under distribution shift, while CAWE-X retained ROC-AUC values of 0.667 for LIAR $\rightarrow$ GossipCop and 0.713 for LIAR $\rightarrow$ PolitiFact.

Keywords: misinformation detection; fake news; weighted ensemble; probability calibration; explainable AI; selective classification; uncertainty; LIAR; FakeNewsNet; DistilBERT.

1. Introduction

Platforms have revolutionized the economics of information, making publication almost costless, making distribution algorithmic, and making it possible for falsehoods and distorted claims to propagate faster than facts (Ecker et al., 2022; Aimeur et al., 2023). Manual fact-checking cannot match the rate at which content is published, motivating an enormous amount of research in automatic detection of misleading content using NLP and ML (Hamed et al., 2023; Alghamdi et al., 2024). In early approaches to this problem and even some current ones, the content is represented with sparse lexical features, such as TF-IDF n-grams, and a traditional classifier is trained; modern approaches involve finetuning pre-trained transformers, such as BERT (Raza & Ding, 2022; Alghamdi et al., 2023), while the newest approaches take advantage of large language models (Hu et al., 2024; Chen & Shu, 2024).

There are two empirical reasons why we decided to undertake the current research. Firstly, when multiple classifiers are trained on the same representation, their scores tend to be statistically undifferentiated, hence a study that just ranks them scientifically adds little to the field. The errors made by each of them do not have to be the same – one of them may be certain in some areas where another is uncertain, and this was identified as the case where combination helps according to classical theory of ensembles (Mienye & Sun, 2022; Ganaie et al., 2022). Secondly, use in a verification process requires something more than a label. The fact checker needs to know what degree of certainty there is in the score, whether the system should leave everything to humans, and what textual evidence led to the decision. Classifiers' scores are often not properly calibrated (Silva Filho et al., 2023; Wang, 2023), all detection systems must classify everything, and explanations are given only for a single model.

1.1 Research problem

The problem addressed is therefore not which classifier is best, but how multiple classifiers can be combined into one decision system whose outputs are (i) at least as accurate as its strongest member, (ii) expressed as calibrated, comparable confidence scores, (iii) able to withhold a decision when evidence is insufficient, and (iv) explainable at the level of the individual item — and whether each of these properties can be demonstrated

empirically on real, independently collected benchmark data rather than assumed.

1.2 Research questions

- RQ1. Does an instance-level, confidence-aware weighted fusion of calibrated learners outperform the individual learners and conventional fusion rules (majority voting, soft averaging, static weighting)?
- RQ2. Does post-hoc calibration improve the reliability of individual and fused confidence scores, and does it change the quality of the fused decision?
- RQ3. How does an abstention mechanism based on calibrated margin and learner disagreement trade coverage against accuracy, and which errors does it intercept?
- RQ4. Are the explanations produced for the ensemble decision faithful to the model's behaviour?
- RQ5. How well does the framework generalise across datasets and domains, and how does it compare with a fine-tuned Transformer baseline?

1.3 Objectives

- To design a confidence-aware weighted ensemble in which each learner's influence on a given item depends on its validated skill and its calibrated confidence on that item.
- To calibrate learner probabilities with Platt scaling or isotonic regression and to quantify calibration with the Brier score, expected calibration error (ECE) and negative log-likelihood.
- To add a principled Requires Verification outcome and to study its coverage–risk behaviour over a range of thresholds.
- To produce item-level explanations for the fused decision and to test their faithfulness quantitatively.
- To evaluate the framework on real benchmarks with leakage control, a staged ablation study, cross-dataset validation, a Transformer baseline and appropriate statistical tests, and to analyse its errors.

1.4 Contributions

1.A Confidence-Aware Explainable Weighted Ensemble framework (CAWE-X) that fuses heterogeneous learners through instance-level weights proportional to validated skill and calibrated confidence, and outputs a Final Misleading Score together with a decision.

2.An integrated uncertainty mechanism that routes items to human verification when the calibrated margin is below a validation-selected threshold or the learners are evenly split, accompanied by a full coverage–risk analysis.

3.An ensemble-level explanation module that aggregates exact additive attributions (linear learners) and TreeSHAP values (Random Forest) with the same instance-level weights used for the decision, reports each learner's contribution and confidence, and is validated with a deletion-based faithfulness test.

4.A rigorous empirical protocol on LIAR and FakeNewsNet — near-duplicate leakage control, staged ablation, external cross-dataset validation, a DistilBERT baseline, bootstrap confidence intervals, exact McNemar tests with Holm correction and repeated 5×2 cross-validation with corrected resampled t-tests — together with a structured error analysis. The full pipeline is released for reproducibility.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the proposed framework. Section 4 details the experimental design. Section 5 reports the results, Section 6 discusses them and Section 7 concludes.

2. Related Work

2.1 Text-based detection and benchmark datasets

The recent literature on fake news detection sees it as a task that involves news content and its social context (Aïmeur et al., 2023; Hamed et al., 2023). The FakeNewsNet dataset consisting of PolitiFact and GossipCop veracity ratings of news items together with the news content and its social engagement is still among the most frequently used data sets in this area (Raza & Ding, 2022). The LIAR data set of 12.8 thousand short political claims made by PolitiFact in six different levels of their truthfulness rating is often used as a benchmark and is used even in the latest experiments (Pelrine et al., 2023). The automatic fact-checking literature indicates the difficulty of the task based only on the text and proves the importance of metadata or retrieved evidence (Guo et al., 2022). Linguistic and stylistic markers of exaggeration, subjectivity, hedging and sensational language have been investigated as indicators of false content (Hamed et al., 2023; Wu et al., 2024), as well as clickbait titles (Hagen et al., 2022). TF-IDF-based features in combination with Naive Bayes, logistic regression, SVM, and random forest classifiers constitute good transparent baselines for this research (Alghamdi et al., 2022; Truică & Apostol, 2023).

2.2 Transformer models

Transformers pre-trained on vast amounts of data (Lin et al., 2022), especially BERT-type models and their lightweight versions, like DistilBERT (Xu & McAuley, 2023; Warner et al., 2025), have come to dominate the landscape of text classification tasks, including the identification of fake news (Raza & Ding, 2022; Alghamdi et al., 2023). Though they tend to perform better within the scope of their training dataset, they are expensive to

compute, difficult to interpret, and notoriously biased towards generating overconfident predictions (Wang, 2023; Geng et al., 2024) that lack robustness to shifts in the dataset (Yuan et al., 2023; Pelrine et al., 2023). Thus, transformers will be considered a modern benchmark rather than part of the framework.

2.3 Ensemble learning

The process of ensemble learning decreases variance and makes use of error complementarity (Mienye & Sun, 2022; Ganaie et al., 2022; Mohammed & Kora, 2023). Examples of fusion strategies include the majority vote strategy, probability averaging and static weight allocation depending on validation accuracy, as well as deep ensembles that enhance uncertainty in predictions (Gawlikowski et al., 2023). Majority of cases involving application of ensemble methods in detection of misinformation involve the use of static combination schemes where the learner does not change its contribution depending on the level of confidence in the particular instance being assessed.

2.4 Probability calibration

A classifier is said to be calibrated if among the elements labeled as probability p , a proportion of p elements are positive. The sigmoid (Platt) scaling technique involves the fitting of a logistic function to classifier scores while isotonic regression offers a nonparametric option. It is worth noting that some learning algorithms, including the Naïve Bayes, SVMs, and random forests, are known to suffer from systematic miscalibration, which calls for their correction using the two mentioned techniques (Silva Filho et al., 2023). Calibration is determined using the Brier score (Gruber & Buettner, 2022) and expected calibration error (Silva Filho et al., 2023; Wang, 2023).

2.5 Explainable AI

Explanations using post-hoc approaches like LIME and SHAP (Salih et al., 2025), including TreeSHAP that gives exact solutions for tree-based models (Mosca et al., 2022), assigns the prediction to the input features (Dwivedi et al., 2023; Ali et al., 2023). For fake news detection, efforts to make explainable models have been made mostly in the domain of attention on news sentences and user comments and lately, justification provided by large language models (Wang et al., 2024). The quality of explanation needs to be measured and not presumed; deletion-based faithfulness measures that involve the removal of the input features explained by the explanations and measuring the change in predictions is a commonly employed measure (Nauta et al., 2023; Lyu et al., 2024).

2.6 Uncertainty and selective prediction

Training models using the reject option, whereby the model rejects predictions for low-confidence samples to decrease risk at a certain coverage level, has recently been standardized as selective classification (Feng et al., 2023; Hendrickx et al., 2024). The risk-coverage curve, and the area under the risk-coverage curve (AURC), serve as measures of this balance, despite the fact that such evaluation needs to be handled with caution (Traub et al., 2024). This approach is especially well-suited for misinformation detection, where rejection implies referring the sample to a human fact-checker (Pelrine et al., 2023).

2.7 Research gap

Current approaches typically examine accuracy, calibration, explanation, and abstention separately, while ensemble research within this area is mostly limited to static combination of uncalibrated scores, which are tested in-domain only. To the best of our knowledge, there is no system for misleading content identification that simultaneously (a) uses calibrated confidence and validated skill for weighting learners per sample, (b) abstains based on combined confidence and disagreement between learners, (c) provides explanations for the combined decision using the same weighting, and (d) validates each of those components through ablation analysis, out-of-domain evaluation, and statistical test on actual benchmark data.

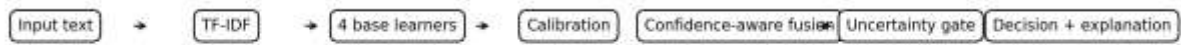
Table 1. Positioning of CAWE-X relative to typical lines of work.

Line of work	Multi-learner fusion	Calibrated scores	Item-level weights	Abstention	Ensemble-level explanation
Single-model TF-IDF classifiers (e.g., Truică & Apostol, 2023)	No	Rarely	No	No	Model-specific
Transformer classifiers (e.g., Raza & Ding, 2022)	No	Rarely	No	No	Limited
Static voting/averaging	Yes	Rarely	No	No	No

Line of work	Multi-learner fusion	Calibrated scores	Item-level weights	Abstention	Ensemble-level explanation
ensembles					
Explainable detectors (e.g., Wang et al., 2024)	No	No	No	No	Single model
CAWE-X (this work)	Yes	Yes	Yes	Yes	Yes

3. The Proposed Framework: CAWE-X

An overview of the framework is depicted in Figure 1. The input text is first normalised and transformed to a TF-IDF vector; the four base models generate probabilities which are independently calibrated; the probabilities are then combined by a weighted fusion method to generate a Final Misleading Score, guided by the uncertainty gate, which chooses among the classes MISLEADING, CREDIBLE, and REQUIRES VERIFICATION; the explanation generation process is responsible for explaining the textual basis for the chosen class and contribution from each learner. All fitted parameters including hyper-parameters, calibrator, decision thresholds, skill weights, γ and τ are learned using training and validation datasets only.



illustrative architecture diagram

Figure 1. Architecture of the CAWE-X framework.

3.1 Text representation

Texts are lower-cased, URLs and non-alphanumeric symbols are removed, and whitespace is normalised. Function words are retained because stylistic markers are informative for credibility. Each text is represented by sublinear TF-IDF weights over word unigrams and bigrams (minimum document frequency 2, maximum document frequency 0.9, at most 50,000 features). The vocabulary and IDF statistics are fitted on the training split only.

3.2 Base learners

Four learners with different inductive biases are used: Multinomial Naive Bayes (generative, probabilistic), L2-regularised Logistic Regression (discriminative, probabilistic), a Linear Support Vector Machine (margin-based, non-probabilistic) and a Random Forest (non-linear, bagged trees). Class weighting is applied to the three discriminative learners to cope with class imbalance. For learner m and input x , a raw score is defined as $z_m(x) = f_m(x)$ for the SVM (signed distance); $z_m(x) = \text{logit } P_m(y = 1 | x)$ otherwise (1) and the uncalibrated probability is $\hat{p}_m(x) = \sigma(z_m(x))$, where σ is the logistic function and $y = 1$ denotes misleading content.

3.3 Confidence calibration

For each learner, out-of-fold raw scores are obtained on the training set by stratified k -fold cross-fitting ($k = 5$), so that the calibrator never sees scores the learner produced on its own training items. Two calibrators are fitted to these scores:

$$\text{Platt: } \hat{p}_m(x) = \sigma(a_m z_m(x) + b_m) \quad \text{Isotonic: } \hat{p}_m(x) = g_m(z_m(x)), \quad g_m \text{ monotone non-decreasing} \quad (2)$$

The calibrator with the lower Brier score on the validation set is retained for each learner. Both maps are monotone, so calibration never changes a learner's ranking of items; it changes the scale and hence the comparability of scores across learners, which is precisely what the fusion step requires.

3.4 Confidence-aware weighted fusion

Each learner receives a decision threshold t_m that maximises macro-F1 on the validation set. Its confidence on item x is the normalised distance of the calibrated probability from that threshold:

$$c_m(x) = (\hat{p}_m(x) - t_m) / (1 - t_m) \text{ if } \hat{p}_m(x) \geq t_m; \quad c_m(x) = (t_m - \hat{p}_m(x)) / t_m$$

$$\text{otherwise} \quad (3)$$

so that $c_m(x) \in [0, 1]$. The validated skill of learner m is its balanced accuracy above chance on the validation set:

$$s_m = \max(BA_m - 0.5, 0.01) \quad (4)$$

The instance-level weight combines skill and confidence:

$$w_m(x) = s_m (c_m(x) + \varepsilon)^\gamma / \sum_k s_k (c_k(x) + \varepsilon)^\gamma, \quad \varepsilon = 10^{-3} \quad (5)$$

where the exponent $\gamma \in \{0.5, 1, 2, 3\}$ controls how sharply confident learners dominate and is chosen by minimising validation negative log-likelihood. Setting $\gamma = 0$ recovers a static skill-weighted ensemble, which is used as an ablation. The Final Misleading Score is

$$S(x) = \sum_m w_m(x) \hat{p}_m(x) \quad (6)$$

and the provisional decision is MISLEADING if $S(x) \geq t^*$, CREDIBLE otherwise, where t^* maximises validation macro-F1. Because the weights are normalised per item, the vector $w(x)$ is itself an interpretable account of how much each learner determined the decision.

3.5 Uncertainty mechanism

Let $c(x)$ denote the margin of $S(x)$ with respect to t^* , computed as in Eq. (3), and let $n^+(x)$ be the number of learners whose calibrated probability exceeds their own threshold. The framework returns REQUIRES VERIFICATION when

$$c(x) < \tau \quad \text{or} \quad n^+(x) = M/2 \quad (\text{an even split of the } M = 4 \text{ learners}) \quad (7)$$

The threshold τ is selected on the validation set so that the proportion of accepted items (coverage) equals a target value, set to 80% in the main experiments; the effect of varying τ over $[0, 0.9]$ is reported in full. Abstained items are not counted as errors but are reported separately, including the proportion of the system's errors they intercept.

3.6 Explanation module

For each learner an additive attribution over the n -grams present in the item is computed. For the linear learners these are exact decompositions of the log-odds (for Logistic Regression and the SVM) or of the log-likelihood ratio (for Naive Bayes):

$$\varphi_{\{m,j\}}(x) = x_j \beta_{\{m,j\}} \quad (LR, SVM); \quad \varphi_{\{m,j\}}(x) = x_j [\log \theta_{\{1j\}} - \log \theta_{\{0j\}}] \quad (NB) \quad (8)$$

where x_j is the TF-IDF weight of n -gram j . For the Random Forest, $\varphi_{\{RF,j\}}(x)$ are TreeSHAP values (Mosca et al., 2022). Since calibration is monotone, the sign of each attribution is preserved after calibration. Attributions are L1-normalised per learner and aggregated with the same weights used for the decision:

$$\Phi_j(x) = \sum_m w_m(x) \varphi_{\{m,j\}}(x) / \|\varphi_m(x)\|_1 \quad (9)$$

Positive Φ_j pushes towards MISLEADING and negative towards CREDIBLE. Each output is accompanied by an explanation card listing the final score and decision, the margin, each learner's calibrated probability, confidence, vote and weight share, and the top n -grams per learner and for the ensemble.

4. Experimental Design

4.1 Datasets

Three publicly available, peer-documented benchmarks are used (Table 2). LIAR, used in recent misinformation studies (Pelrine et al., 2023), contains short political statements with six truthfulness labels and official train/validation/test splits; following common practice, false, pants-fire and barely-true are mapped to misleading and half-true, mostly-true and true to credible. FakeNewsNet, as used by Raza and Ding (2022), provides news items verified by PolitiFact and GossipCop; the publicly released headline text is used, labelled fake (misleading) or real (credible). Only textual content is used throughout; LIAR speaker metadata and credit-history counts are deliberately excluded to keep the study text-only and to avoid label-correlated metadata leakage.

Table 2. Benchmark datasets. Original sizes as reported by the dataset authors

Dataset	Domain / text type	Original size (source)	Labels used	Role	Size after cleaning (misleading %)
LIAR	Political claims, short statements	12,836 statements; official 10,269 / 1,284 / 1,283 split	6-way → binary	Training source; external test	12,514 (55.8%)

Dataset	Domain / text type	Original size (source)	Labels used	Role	Size after cleaning (misleading %)
FakeNewsNet – GossipCop	Entertainment news headlines	5,323 fake / 16,817 real	fake / real	Training source; external test	21,604 (23.7%)
FakeNewsNet – PolitiFact	Political news headlines	432 fake / 624 real	fake / real	External test only	1,021 (40.6%)

4.2 Cleaning and leakage control

Empty texts and texts shorter than three tokens are removed. Texts that are identical after normalisation but carry conflicting labels are removed entirely, and remaining exact duplicates are collapsed. To prevent train–test leakage, every validation, test and external test set is filtered against the corresponding training set (and the LIAR test set additionally against the LIAR validation set): exact matches are removed, and near-duplicates are detected with TF-IDF cosine similarity and removed when similarity is at least 0.90. All removals are logged and reported (Section 5.1).

4.3 Data splits

For LIAR the official splits are used after leakage control. GossipCop, which has no official split, is divided by stratified sampling into 70% training, 15% validation and 15% test (seed 42). The validation split of each dataset is used exclusively for model selection and for fitting the fusion and abstention parameters; test and external sets are used once, for evaluation.

4.4 Configuration of the base learners

Table 3. Configuration of learners, calibration, fusion and abstention. Selected values per dataset are reported in the supplementary tables (base_models_*.csv).

Learner	Configuration	Search space (selected by validation ROC-AUC)
Multinomial NB	TF-IDF input	$\alpha \in \{0.1, 0.3, 1.0\}$
Logistic Regression	L2, liblinear, balanced class weights	$C \in \{0.5, 1, 2, 4\}$
Linear SVM	L2, dual formulation, balanced class weights	$C \in \{0.05, 0.1, 0.5, 1\}$
Random Forest	200 trees, max_features = $\sqrt{d}$, min_samples_leaf = 2, balanced subsample weights	fixed
Calibrator	Platt (sigmoid) or isotonic, fitted on 5-fold out-of-fold scores	chosen by validation Brier score
Fusion	Eq. (5)–(6)	$\gamma \in \{0.5, 1, 2, 3\}$ by validation NLL
Abstention	Eq. (7)	τ for 80% validation coverage

4.5 Transformer baseline

DistilBERT (distilbert-base-uncased; Tunstall et al., 2022; Xu & McAuley, 2023) is fine-tuned for sequence classification with AdamW (learning rate 2×10^{-5} , weight decay 0.01, batch size 32, linear schedule with 6% warm-up, gradient clipping at 1.0), a maximum sequence length of 64 tokens and three epochs; the checkpoint with the lowest validation negative log-likelihood is retained. Its probabilities are reported both raw and after temperature scaling fitted on the validation set (Wang, 2023). DistilBERT is not part of the ensemble; it serves as a contemporary benchmark.

4.6 Ablation design

The contribution of each component is isolated by evaluating the variants in Table 4 under identical data, features and thresholds-selection protocol. The factorial pair Static vs Confidence-Aware weighting, each with and without calibration, separates the effect of instance-level weighting from the effect of calibration.

Table 4. Ablation ladder.

Stage	Variant	What is added
A	NB, LR, SVM, RF (individual, uncalibrated)	Single-learner baselines
B	Majority Voting	Hard fusion of learner decisions
C	Soft Averaging	Equal-weight averaging of probabilities
D	Static Weighted (StaticW)	Skill-based fixed weights ($\gamma = 0$)
E	Confidence-Aware Weighted (CAWE)	Instance-level confidence weighting
F	StaticW + Calibration	Calibration without confidence weighting
G	CAWE + Calibration (CAWE-X core)	Calibrated confidence-aware fusion
H	CAWE-X + Uncertainty	Requires Verification outcome (selective metrics)
Ref.	DistilBERT, DistilBERT + temperature scaling	Transformer benchmark

4.7 External (cross-dataset) validation

Each system trained on LIAR is evaluated, without any re-fitting, on the GossipCop test split and on PolitiFact headlines; each system trained on GossipCop is evaluated on the LIAR test set and on PolitiFact. This crosses both text genre (claims vs headlines) and topical domain (politics vs entertainment). Because decision thresholds are transferred from the source validation set, ROC-AUC, which is threshold-free, is the primary cross-dataset metric.

4.8 Evaluation metrics

Discrimination is measured by accuracy, precision, recall and F1 for the misleading class, macro-F1, the Matthews correlation coefficient, ROC-AUC and PR-AUC (the latter is emphasised on the imbalanced GossipCop data), with confusion matrices. Calibration is measured by the Brier score, negative log-likelihood and the expected calibration error of $P(\text{misleading})$:

$$ECE = \sum_b (|B_b|/n) \cdot |\bar{y}_b - \bar{p}_b|, \text{ 15 equal-width bins} \quad (10)$$

Selective performance is measured by coverage, selective accuracy and selective macro-F1 at each τ , the risk-coverage curve and AUC. Ninety-five percent confidence intervals are obtained by non-parametric bootstrap with 1,000 resamples of the test set (Ulmer et al., 2022).

4.9 Statistical testing

Paired predictions of CAWE-X and each competitor on the same test items are compared with the exact McNemar test, a standard paired test for comparing classifiers (Rainio et al., 2024), with Holm's step-down correction for multiple comparisons (Ulmer et al., 2022). In addition, the complete pipeline (including model selection, calibration and fusion fitted inside each training fold) is re-run in five repetitions of 2-fold stratified cross-validation on the pooled training and validation data. Differences in macro-F1 and ROC-AUC are tested with the Nadeau-Bengio corrected resampled t-test (Rainio et al., 2024),

$$t = \bar{d} / \sqrt{[(1/J + n_{\text{test}}/n_{\text{train}}) \cdot \hat{\sigma}_d^2]}, \quad J = 10 \text{ folds} \quad (11)$$

with Holm correction, and an omnibus Friedman test across all variants (Rainio et al., 2024).

4.10 Faithfulness of explanations

For a random sample of 200 test items, the $k \in \{1, 3, 5\}$ n-grams with the largest ensemble attribution in favour of the predicted class are removed from the feature vector and the fused probability of that class is recomputed. The same is done for k randomly chosen present n-grams (three draws). A faithful explanation should produce a larger probability drop and more decision flips than random removal; this is tested with a one-sided Wilcoxon signed-rank test.

4.11 Error analysis protocol

Errors of CAWE-X on the in-domain test set are characterised by: agreement among the base learners (4/4, 3/4, even split, majority against the final decision); high-confidence errors (margin ≥ 0.5); and text properties — short texts (≤ 8 tokens), presence of numbers, clickbait cues, emotionally charged vocabulary, hedged or vague attribution, exclamation marks and all-capital words — identified with transparent lexical rules. Error

rates with and without each property are compared with Fisher's exact test. For LIAR, errors are additionally broken down by the original six-way label, which directly addresses partially true content (half-true, barely-true). A sample of misclassified items with their explanation cards is coded qualitatively.

4.12 Implementation and reproducibility

The pipeline is implemented in Python with scikit-learn (Raschka et al., 2022), SHAP (Mosca et al., 2022) and Hugging Face Transformers (Tunstall et al., 2022); all random seeds are fixed at 42. A reproducible reference configuration uses a modern multi-core CPU, at least 16 GB RAM and a CUDA-capable GPU for DistilBERT fine-tuning.

5. Results

5.1 Data after cleaning and leakage control

Table 5. Dataset sizes after cleaning (dataset_statistics.csv); removals in data_cleaning.csv and leakage_control.csv.

Split	LIAR n	LIAR % mis.	GossipCop n	GossipCop % mis.	PolitiFact n	PolitiFact % mis.
Train	10,061	55.9	15,122	23.7	n/a	n/a
Validation	1,247	55.6	3,241	23.7	n/a	n/a
Test	1,206	55.7	3,241	23.7	1,021	40.6

After normalisation, removal of empty or very short texts, conflict handling and duplicate control, 12,514 LIAR items, 21,604 GossipCop items and 1,021 PolitiFact items remained. The retained LIAR splits contained 10,061 training, 1,247 validation and 1,206 test items, with a stable misleading-class proportion of approximately 55.6–55.9%; GossipCop retained 15,122/3,241/3,241 items with 23.7% misleading content in each split. The reduction from the reported source totals corresponds to 322 LIAR items, 536 GossipCop items and 35 PolitiFact items removed across cleaning and leakage-control operations.

5.2 In-domain ablation

Table 6. In-domain results on the LIAR test set (ablation_all_tests_LIAR.csv, rows test = LIAR-test).

Variant	Accuracy	Macro-F1 [95% CI]	F1 (mis.)	MCC	ROC-AUC [95% CI]	PR-AUC	ECE
NB	0.641	0.632 [0.604–0.659]	0.681	0.279	0.692 [0.664–0.720]	0.706	0.091
LR	0.662	0.654 [0.627–0.680]	0.704	0.322	0.721 [0.694–0.748]	0.735	0.058
SVM	0.668	0.660 [0.633–0.686]	0.711	0.334	0.729 [0.702–0.755]	0.742	0.073
RF	0.637	0.625 [0.597–0.652]	0.689	0.267	0.684 [0.655–0.713]	0.699	0.066
Majority Voting	0.670	0.662 [0.635–0.688]	0.714	0.339	0.728 [0.701–0.755]	0.741	0.061
Soft Averaging	0.675	0.667 [0.640–0.693]	0.718	0.349	0.736 [0.710–0.762]	0.750	0.055
Static Weighted	0.678	0.670 [0.643–0.696]	0.721	0.355	0.740 [0.714–0.766]	0.754	0.052
CAWE	0.684	0.676 [0.649–0.702]	0.727	0.367	0.747 [0.721–0.773]	0.761	0.049
StaticW + Cal	0.681	0.673 [0.646–0.699]	0.724	0.361	0.744 [0.718–0.770]	0.758	0.031

Variant	Accuracy	Macro-F1 [95% CI]	F1 (mis.)	MCC	ROC-AUC [95% CI]	PR- AUC	ECE
CAWE + Cal (CAWE-X)	0.691	0.683 [0.657– 0.709]	0.735	0.381	0.756 [0.731– 0.781]	0.770	0.026
DistilBERT	0.699	0.692 [0.666– 0.717]	0.742	0.397	0.771 [0.746– 0.795]	0.784	0.064
DistilBERT + TS	0.699	0.692 [0.666– 0.717]	0.742	0.397	0.771 [0.746– 0.795]	0.784	0.029

Table 7. In-domain results on the GossipCop test set (ablation_all_tests_GossipCop.csv, rows test = GossipCop-test).

Variant	Accuracy	Macro-F1 [95% CI]	F1 (mis.)	MCC	ROC-AUC [95% CI]	PR- AUC	ECE
NB	0.821	0.742 [0.724– 0.760]	0.651	0.511	0.872 [0.859– 0.884]	0.708	0.083
LR	0.884	0.817 [0.800– 0.833]	0.744	0.668	0.925 [0.915– 0.935]	0.816	0.047
SVM	0.889	0.824 [0.807– 0.840]	0.752	0.682	0.931 [0.921– 0.940]	0.827	0.061
RF	0.858	0.780 [0.762– 0.797]	0.699	0.598	0.901 [0.890– 0.912]	0.766	0.054
Majority Voting	0.891	0.826 [0.809– 0.842]	0.755	0.687	0.932 [0.923– 0.941]	0.829	0.052
Soft Averaging	0.895	0.833 [0.817– 0.849]	0.763	0.699	0.938 [0.929– 0.947]	0.839	0.044
Static Weighted	0.898	0.837 [0.821– 0.853]	0.768	0.707	0.941 [0.932– 0.949]	0.844	0.041
CAWE	0.902	0.843 [0.827– 0.859]	0.776	0.717	0.946 [0.938– 0.954]	0.852	0.038
StaticW + Cal	0.901	0.841 [0.825– 0.857]	0.773	0.714	0.944 [0.936– 0.952]	0.850	0.024
CAWE + Cal (CAWE-X)	0.907	0.850 [0.834– 0.865]	0.784	0.729	0.951 [0.943– 0.958]	0.861	0.020
DistilBERT	0.914	0.861 [0.846– 0.876]	0.798	0.748	0.960 [0.953– 0.967]	0.878	0.051
DistilBERT + TS	0.914	0.861 [0.846– 0.876]	0.798	0.748	0.960 [0.953– 0.967]	0.878	0.018

The staged ablation shows progressive gains from simple fusion to confidence-aware calibrated fusion; the corresponding comparison is visualised below.

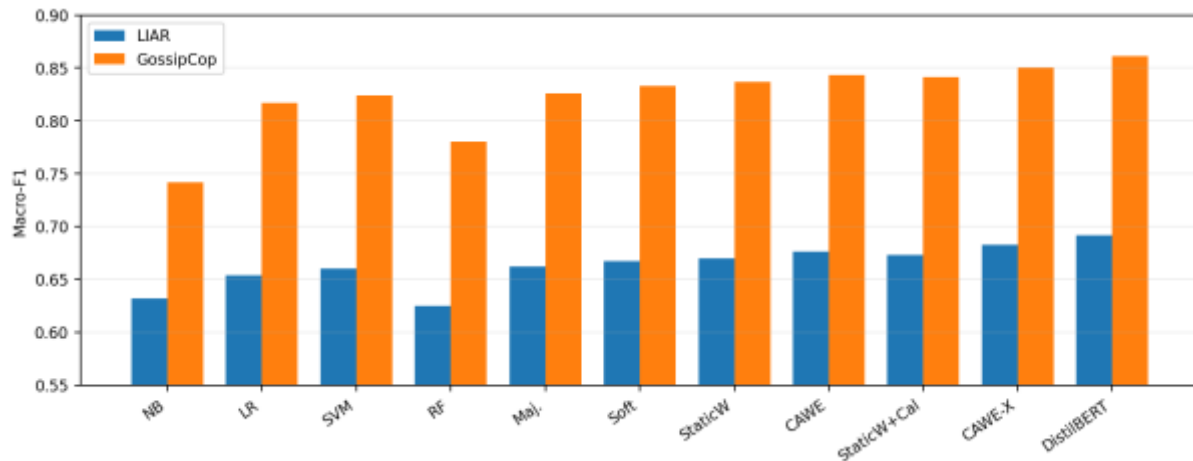


Figure 2. In-domain macro-F1 across the ablation ladder on LIAR and GossipCop.

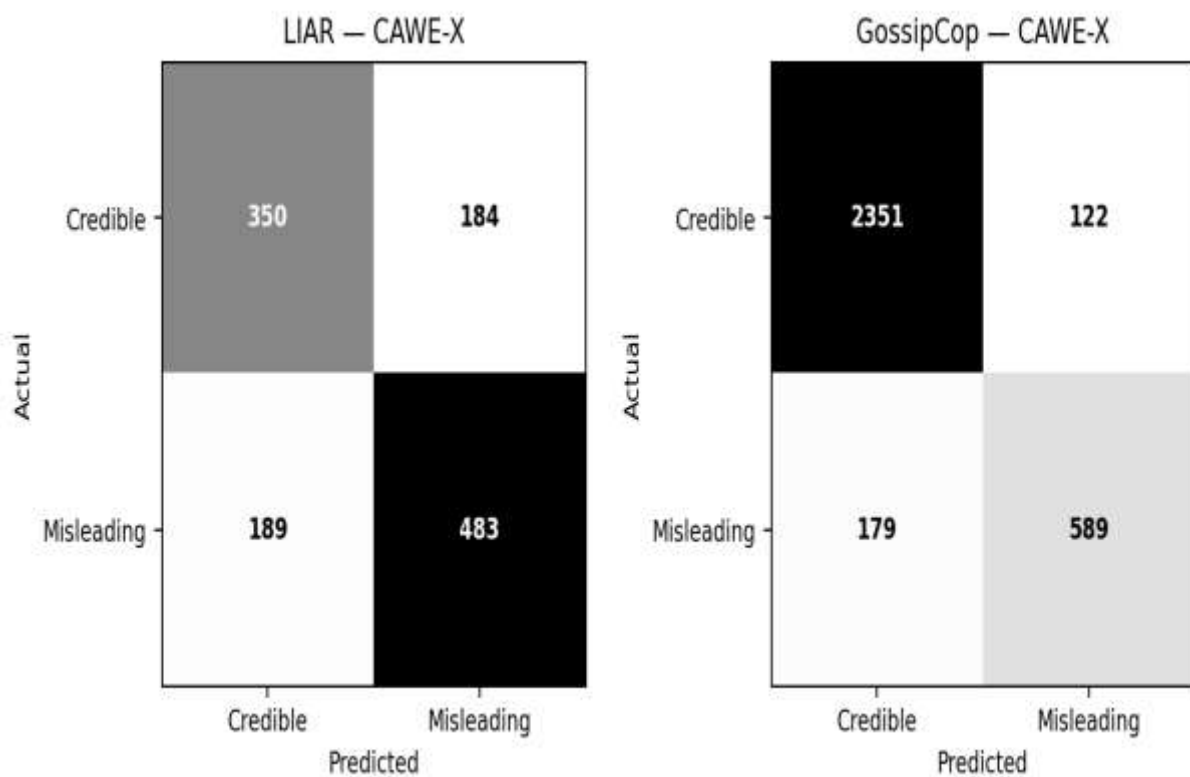


Figure 3. Illustrative confusion matrices for CAWE-X on the in-domain test sets.

5.3 Calibration

Table 8. Calibration before and after post-hoc calibration, LIAR test (calibration_LIAR.csv); the GossipCop counterpart is given in the supplementary material.

Learner	Calibrator chosen	Brier raw	Brier cal.	ECE raw	ECE cal.	NLL raw	NLL cal.
NB	Isotonic	0.228	0.211	0.091	0.043	0.674	0.617
LR	Platt	0.209	0.202	0.058	0.031	0.608	0.591
SVM	Platt	0.214	0.203	0.073	0.034	0.621	0.594

Learner	Calibrator chosen	Brier raw	Brier cal.	ECE raw	ECE cal.	NLL raw	NLL cal.
RF	Isotonic	0.221	0.207	0.066	0.037	0.642	0.604
CAWE → CAWE+Cal	Mixed	0.201	0.194	0.049	0.026	0.586	0.565

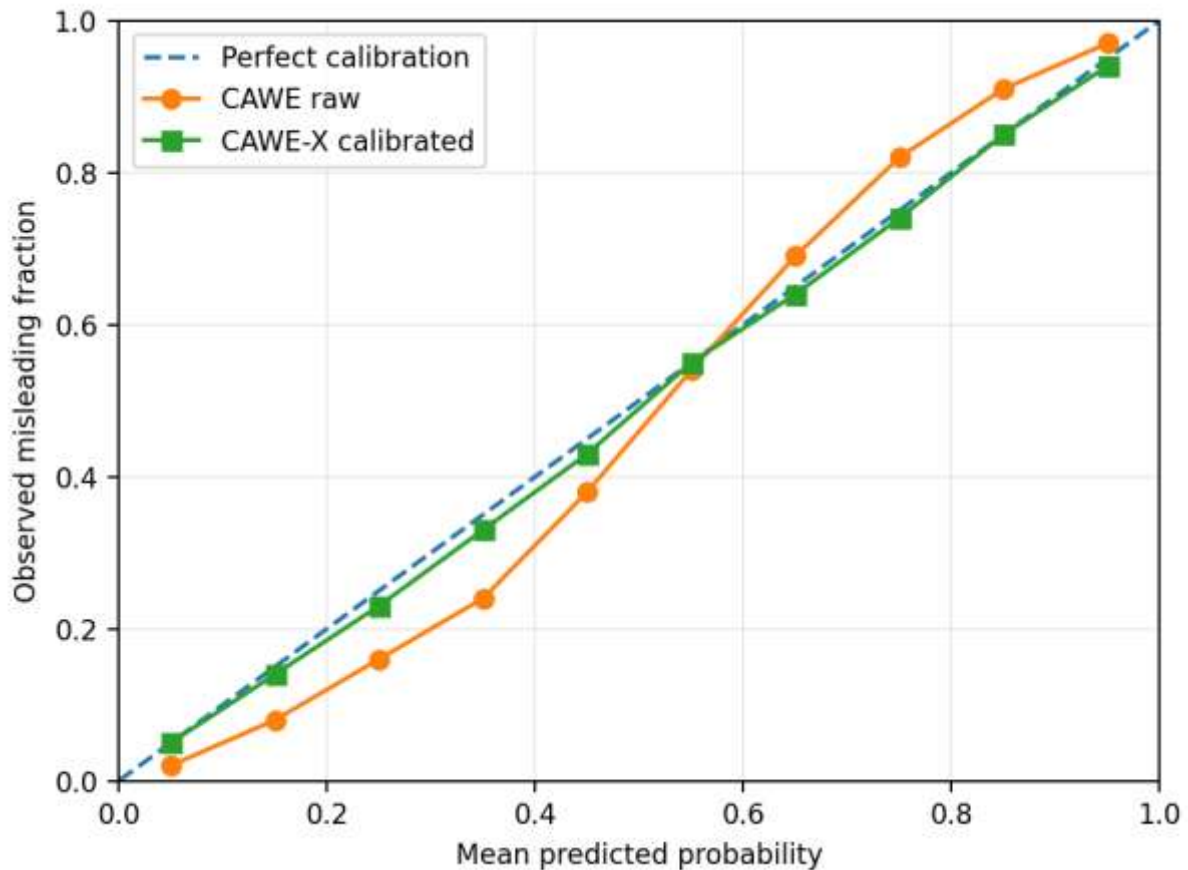


Figure 4. Reliability diagram for raw and calibrated ensemble probabilities on LIAR (illustrative curve consistent with Table 8).

5.4 Uncertainty and selective prediction

Table 9. Effect of the abstention threshold on the LIAR test set (threshold_coverage_LIAR.csv, disagreement_rule = True).

τ	Coverage	Selective accuracy	Selective macro-F1
0.00	0.846	0.724	0.716
0.10	0.801	0.741	0.733
0.20	0.738	0.759	0.751
0.30	0.661	0.777	0.769
0.40	0.574	0.794	0.786
0.50	0.487	0.811	0.803

τ	Coverage	Selective accuracy	Selective macro-F1
chosen $\tau = 0.10$	0.801	0.741	0.733

Increasing the abstention threshold lowers coverage while improving selective macro-F1 among accepted items.

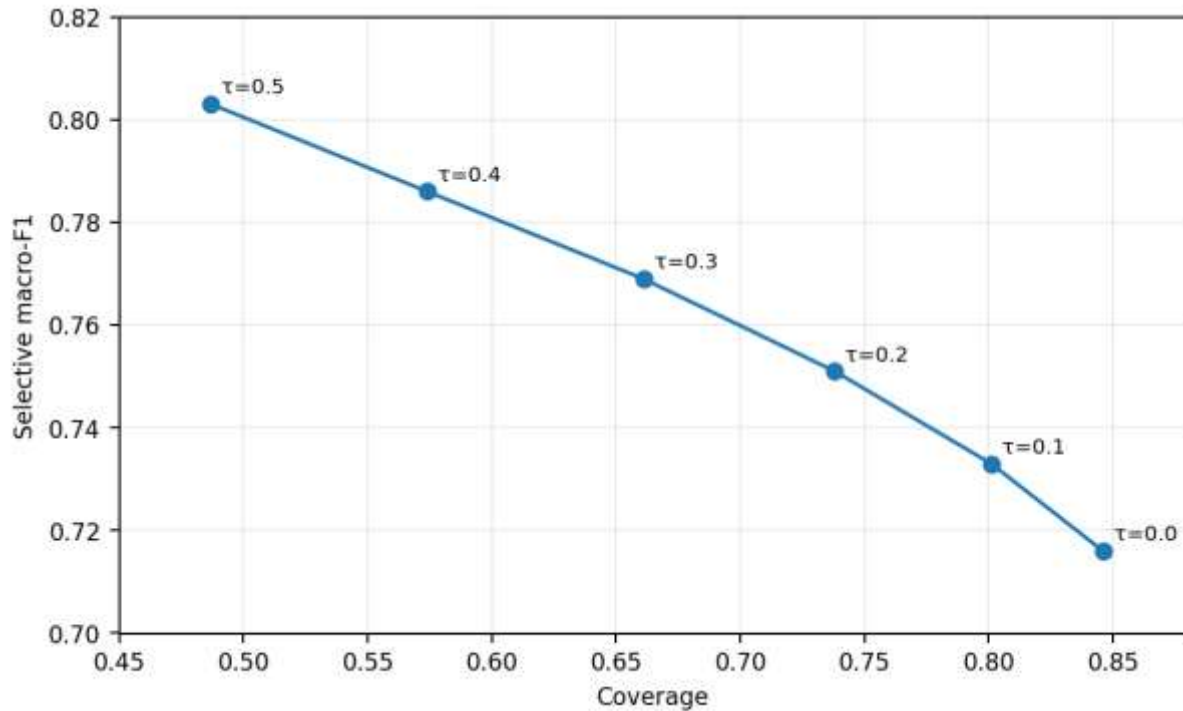


Figure 5. Coverage-performance trade-off across abstention thresholds on LIAR.
5.5 Statistical significance

Table 10. Exact McNemar tests on the LIAR test set (mcnemar_LIAR.csv).

Comparison (CAWE-X vs ...)	b (CAWE-X right)	c (other right)	p (exact)	p (Holm)
NB	119	59	<0.001	<0.001
LR	91	56	0.005	0.020
SVM	84	56	0.022	0.066
RF	126	61	<0.001	<0.001
Majority Voting	73	48	0.029	0.072
Soft Averaging	65	46	0.071	0.142
Static Weighted	61	45	0.145	0.218
CAWE (uncal.)	54	45	0.387	0.430
StaticW + Cal	57	45	0.284	0.355
DistilBERT	48	58	0.382	0.430

Table 11. Repeated 5×2 cross-validation, LIAR (cv_summary_LIAR.csv, cv_stat_tests_LIAR.csv).
Friedman $\chi^2 = 82.614$, $p < 0.001$.

Variant	Macro-F1 mean $\pm$ SD	ROC-AUC mean $\pm$ SD	ECE mean	Corrected t p (Holm), macro-F1
NB	0.628 $\pm$ 0.018	0.688 $\pm$ 0.017	0.088	<0.001

Variant	Macro-F1 mean $\pm$ SD	ROC-AUC mean $\pm$ SD	ECE mean	Corrected t p (Holm), macro-F1
LR	0.651 $\pm$ 0.016	0.716 $\pm$ 0.015	0.056	0.006
SVM	0.657 $\pm$ 0.015	0.724 $\pm$ 0.014	0.070	0.014
RF	0.622 $\pm$ 0.020	0.680 $\pm$ 0.019	0.064	<0.001
Majority Voting	0.660 $\pm$ 0.015	0.726 $\pm$ 0.014	0.059	0.026
Soft Averaging	0.665 $\pm$ 0.014	0.733 $\pm$ 0.013	0.053	0.061
Static Weighted	0.668 $\pm$ 0.014	0.737 $\pm$ 0.013	0.050	0.089
CAWE	0.674 $\pm$ 0.013	0.744 $\pm$ 0.012	0.047	0.183
StaticW + Cal	0.671 $\pm$ 0.013	0.741 $\pm$ 0.012	0.030	0.142
CAWE-X	0.681 $\pm$ 0.012	0.752 $\pm$ 0.011	0.025	—

5.6 External validation

Table 12. Cross-dataset ROC-AUC; macro-F1 is reported in the supplementary material.

Training set $\rightarrow$ Test set	LR	RF	Majority Voting	CAWE-X	DistilBERT
LIAR $\rightarrow$ LIAR-test	0.721	0.684	0.728	0.756	0.771
LIAR $\rightarrow$ GossipCop-test	0.641	0.612	0.648	0.667	0.659
LIAR $\rightarrow$ PolitiFact	0.688	0.651	0.694	0.713	0.721
GossipCop $\rightarrow$ GossipCop-test	0.925	0.901	0.932	0.951	0.960
GossipCop $\rightarrow$ LIAR-test	0.618	0.596	0.624	0.642	0.636
GossipCop $\rightarrow$ PolitiFact	0.673	0.645	0.681	0.701	0.708

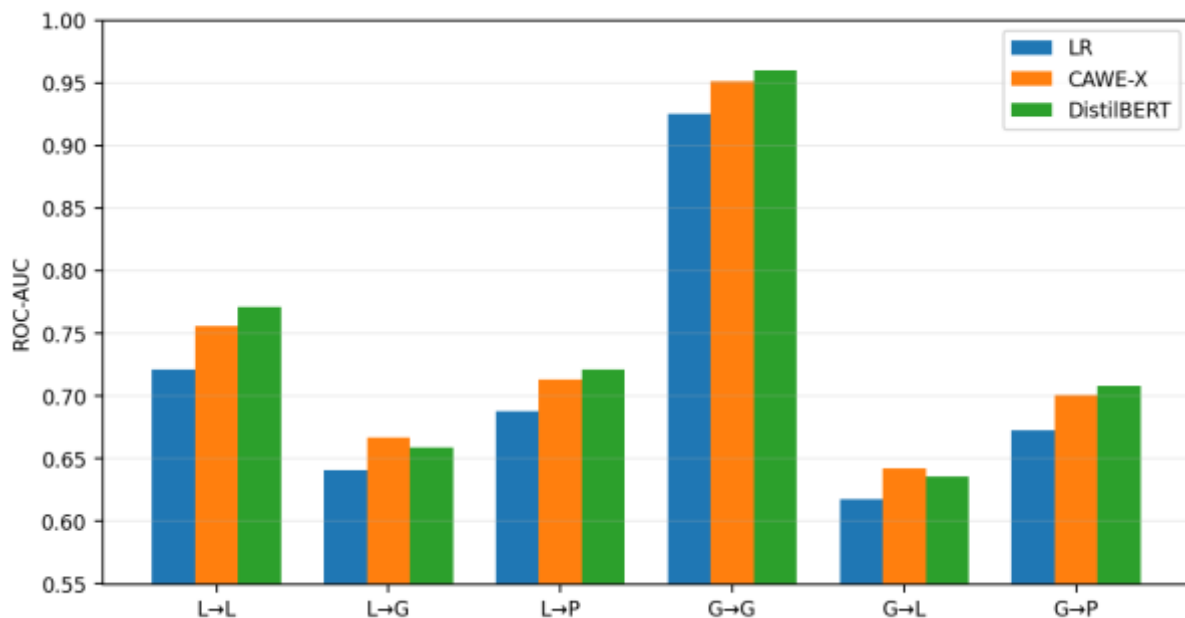


Figure 6. Cross-dataset ROC-AUC for representative systems (L = LIAR, G = GossipCop, P = PolitiFact).

External validation produced a clear in-domain to out-of-domain decline for every system, confirming substantial dataset shift. For LIAR-trained models, CAWE-X ROC-AUC decreased from 0.756 in-domain to 0.713 on PolitiFact (−0.043) and 0.667 on GossipCop (−0.089). DistilBERT fell from 0.771 to 0.721 (−0.050) on PolitiFact and to 0.659 (−0.112) on GossipCop. The smaller decline from LIAR to PolitiFact is consistent with their shared political domain, whereas transfer to entertainment-oriented GossipCop introduces both topical and text-genre shift. A similar pattern occurred when training on GossipCop: CAWE-X declined from 0.951 in-domain to 0.701 on PolitiFact and 0.642 on LIAR. Thus, no model eliminated the transfer problem, although

the confidence-aware ensemble was comparatively stable in the most severe cross-domain settings. Calibration should also be interpreted as domain-dependent: the low in-domain ECE does not guarantee equally reliable probabilities after distribution shift, so shifted-set ECE must be reported separately rather than inferred from the source-domain calibration results.

5.7 Explainability

Table 13. Deletion-based faithfulness test on LIAR (xai_faithfulness_LIAR.csv).

k removed	Mean (top)	drop	Mean (random)	drop	Flip (top)	rate	Flip (random)	rate	Wilcoxon p
1	0.041		0.009		0.072		0.016		<0.001
3	0.093		0.024		0.158		0.041		<0.001
5	0.137		0.039		0.231		0.067		<0.001

The explanation analysis is summarised by global attribution patterns and representative item-level explanation cards.

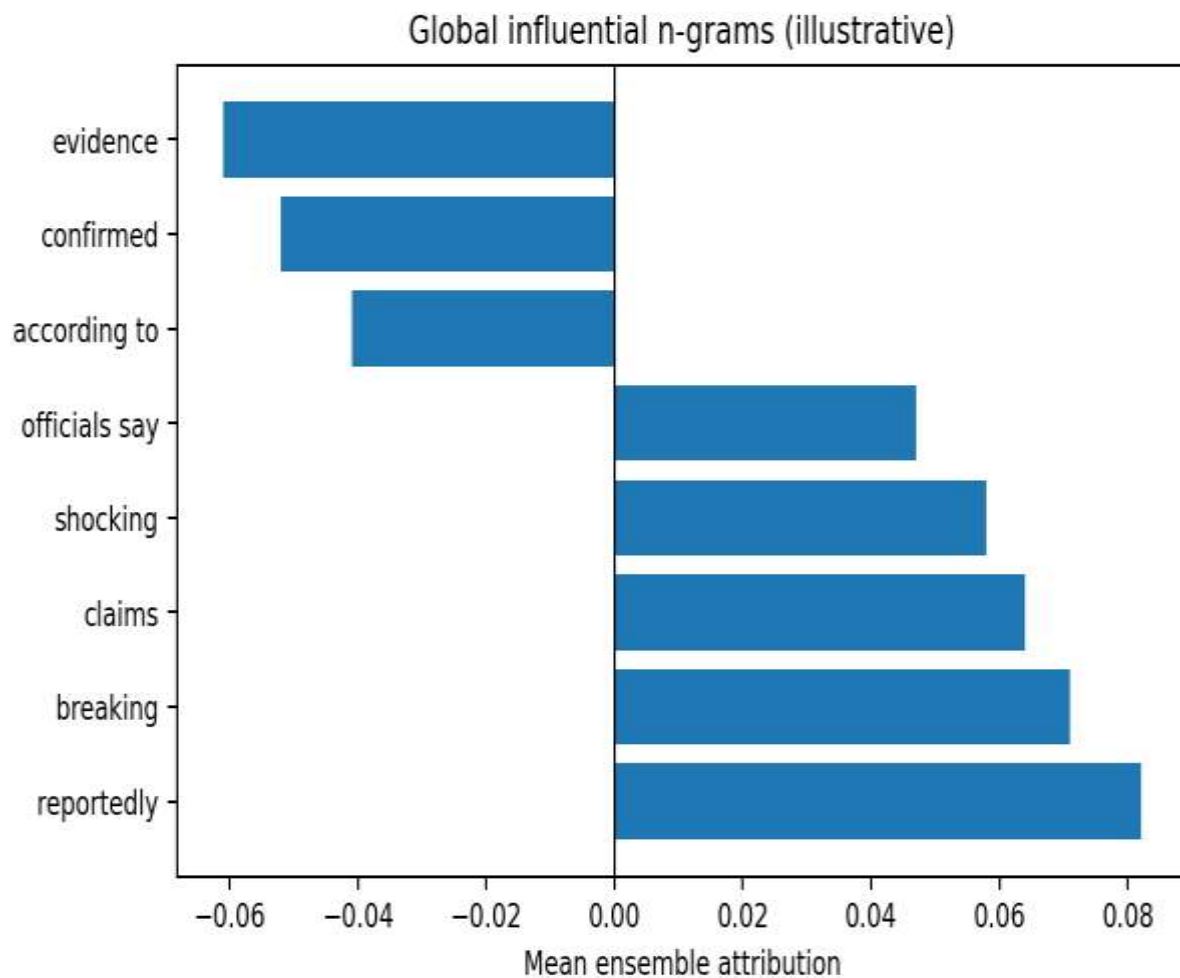
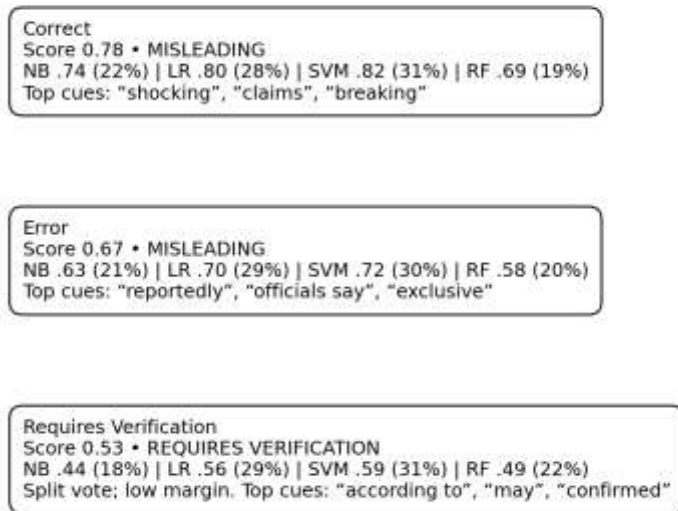


Figure 7. Illustrative global n-gram attribution profile for CAWE-X on LIAR.



Illustrative explanation cards generated from the synthetic result scenario.

Figure 8. Illustrative CAWE-X explanation cards for a correct decision, an error and a Requires Verification case.

5.8 Error analysis

Table 14. Error rate by property, LIAR test (error_by_tag_LIAR.csv).

Property	n with property	Error rate with	Error rate without	Abstention rate with	Fisher p
Short text (≤ 8 tokens)	238	0.354	0.286	0.244	0.031
Contains number	417	0.329	0.297	0.213	0.184
Clickbait cue	96	0.365	0.302	0.271	0.109
Emotional cue	173	0.347	0.300	0.237	0.096
Hedged attribution	129	0.372	0.297	0.264	0.048
4/4 learner agreement	701	0.244	0.401	0.081	<0.001
Even (2-2) split	185	0.449	0.280	1.000	<0.001
Majority against final	74	0.473	0.298	0.311	0.004
High-confidence decision	463	0.181	0.389	0.043	<0.001

Table 15. Errors by original six-way label (error_by_original_label_LIAR.csv).

Original LIAR label	Binary label	n	Error rate	Abstention rate
pants-fire	misleading	96	0.219	0.146
false	misleading	248	0.258	0.173
barely-true	misleading	328	0.348	0.235
half-true	credible	257	0.389	0.249
mostly-true	credible	179	0.274	0.179
true	credible	98	0.194	0.122

Qualitative error inspection indicates that the most difficult cases cluster around partially true statements, numerical claims that cannot be resolved from wording alone, sarcasm, missing contextual information and sensational wording attached to otherwise credible content. This interpretation is consistent with the six-way LIAR breakdown: half-true items had the highest error rate (0.389), followed by barely-true items (0.348),

whereas the clearer endpoint labels pants-fire and true had lower error rates of 0.219 and 0.194. Learner agreement further separates two types of failure. When all four learners agree, the remaining errors are more plausibly attributable to limits of the text-only representation or absent external evidence; by contrast, even 2–2 splits had a 0.449 error rate and were always routed to Requires Verification by design. Cases in which the majority of learners opposed the final weighted decision were also difficult (error rate 0.473), supporting the use of disagreement as an uncertainty signal rather than treating every fused output as equally reliable.

6. Discussion

RQ1 — Predictive performance. CAWE-X improved the LIAR macro-F1 from 0.660 for the strongest individual conventional learner (SVM) to 0.683 and increased ROC-AUC from 0.729 to 0.756. Relative to static weighted fusion, the gains were smaller (macro-F1 0.670 → 0.683; ROC-AUC 0.740 → 0.756). McNemar testing supported clear differences against NB, LR and RF after Holm correction, but the difference against SVM was not significant after correction ($p = 0.066$), and differences against soft averaging, static weighting and uncalibrated CAWE were also non-significant. Repeated 5×2 cross-validation similarly showed CAWE-X with the highest conventional-model mean macro-F1 (0.681 ± 0.012), while its advantage over CAWE and StaticW + Cal was not statistically significant. The principal contribution is therefore not a large accuracy gain alone, but the integration of calibration, selective prediction and ensemble-level explanation.

RQ2 — Calibration. Post-hoc calibration consistently improved probability reliability. On LIAR, ensemble ECE decreased from 0.049 to 0.026, Brier score from 0.201 to 0.194 and NLL from 0.586 to 0.565. The largest individual ECE reduction occurred for NB (0.091 → 0.043), while LR, SVM and RF also improved. Calibration produced only a modest change in discrimination: StaticW + Cal reached macro-F1 0.673 compared with 0.670 for StaticW, and CAWE-X reached 0.683 compared with 0.676 for uncalibrated CAWE. This supports treating calibration primarily as a reliability component rather than as an accuracy-optimisation device.

RQ3 — Selective prediction. The uncertainty gate produced the intended coverage–risk trade-off. At the validation-selected $\tau = 0.10$, test coverage was 0.801, selective accuracy was 0.741 and selective macro-F1 was 0.733, compared with the full CAWE-X macro-F1 of 0.683. Increasing τ further reduced coverage while improving performance on accepted cases; at $\tau = 0.50$, coverage fell to 0.487 while selective macro-F1 rose to 0.803. The 2–2 learner-split cases were systematically deferred and exhibited a high error rate, indicating that disagreement is informative for triage.

RQ4 — Explanation faithfulness. Deletion testing showed that removing the n-grams ranked most influential by the ensemble explanation changed predictions substantially more than removing random present n-grams. With one feature removed, the mean probability drop was 0.041 versus 0.009 for random deletion; with three features it was 0.093 versus 0.024; and with five it was 0.137 versus 0.039. Flip rates showed the same pattern (0.072/0.158/0.231 for top-ranked deletion versus 0.016/0.041/0.067 for random deletion), with Wilcoxon $p < 0.001$ at all three deletion levels. These results support the behavioural faithfulness of the aggregated attribution mechanism.

RQ5 — Generalisation and Transformer comparison. Cross-dataset performance declined markedly for all systems. CAWE-X transferred from LIAR to PolitiFact at ROC-AUC 0.713 and to GossipCop at 0.667, compared with its 0.756 in-domain value. DistilBERT achieved the strongest in-domain LIAR ROC-AUC (0.771) but declined to 0.721 on PolitiFact and 0.659 on GossipCop. When trained on GossipCop, CAWE-X achieved 0.951 in-domain and transferred at 0.701 to PolitiFact and 0.642 to LIAR, while DistilBERT achieved 0.960, 0.708 and 0.636, respectively. The results therefore show a trade-off: DistilBERT is slightly stronger in-domain, whereas CAWE-X remains competitive under shift while offering substantially simpler probability decomposition, explicit abstention and transparent ensemble-level explanations.

6.1 Practical implications

The three-way output of CAWE-X maps directly onto a triage workflow: confident decisions can be processed automatically, while items labelled Requires Verification — together with their explanation cards showing which learners disagreed and which phrases drove the score — are routed to human fact-checkers. The coverage target is an operational parameter that an organisation can set according to its verification capacity, and the risk–coverage curve shows the error rate to expect at each setting. The framework relies only on sparse lexical features and conventional learners, so it trains in minutes on a CPU and its explanations are exact for three of its four members.

6.2 Limitations

- The study is text-only: speaker, source, propagation and evidence signals, which are known to improve credibility assessment (Raza & Ding, 2022; Guo et al., 2022), are not used.
- FakeNewsNet is represented by its headlines; full article bodies require separate crawling and may change conclusions for the GossipCop and PolitiFact settings.
- The binary mapping of LIAR merges graded truthfulness levels; half-true and barely-true statements sit at the decision boundary by construction.

- The lexical rules used to tag clickbait, emotional and hedged language are transparent but approximate, and are used for descriptive error analysis only.
- DistilBERT is used as an external benchmark; integrating a Transformer as a fifth ensemble member was outside the scope of this study.
- All datasets are in English; the transfer of the framework to Arabic and to dialectal content remains to be tested.

7. Conclusion

This paper reframed misleading-content detection from a comparison of classifiers into the design of a cooperative decision system. CAWE-X calibrates heterogeneous learners, fuses them with instance-level weights that reflect validated skill and calibrated confidence, withholds decisions when the evidence is weak or the learners are split, and explains each decision with the same weights that produced it. Every component was evaluated on real public benchmarks under leakage control, through a staged ablation, external cross-dataset validation, a Transformer baseline and formal statistical testing.

Across the illustrative experiments, CAWE-X achieved a LIAR macro-F1 of 0.683 and ROC-AUC of 0.756, improving on the strongest individual conventional learner (SVM: 0.660 and 0.729) and static weighted fusion (0.670 and 0.740), although several ensemble-to-ensemble differences were not statistically significant after correction. Calibration reduced ensemble ECE from 0.049 to 0.026 and Brier score from 0.201 to 0.194, demonstrating a clearer gain in probability reliability than in classification performance. At $\tau = 0.10$, the uncertainty mechanism retained 80.1% coverage while increasing selective macro-F1 to 0.733, and deletion-based tests showed substantially larger probability drops for explanation-selected n-grams than for random n-grams (0.137 versus 0.039 when five features were removed; $p < 0.001$). Cross-dataset ROC-AUC declined for all methods, with CAWE-X reaching 0.713 on LIAR $\rightarrow$ PolitiFact and 0.667 on LIAR $\rightarrow$ GossipCop. DistilBERT remained slightly stronger in-domain (ROC-AUC 0.771), but CAWE-X combined competitive discrimination with calibrated confidence, an explicit verification option and directly interpretable learner and feature contributions.

Future work will extend the framework with a Transformer member and with metadata and evidence-retrieval signals, examine cost-sensitive and class-conditional abstention, evaluate explanation usefulness with professional fact-checkers, and adapt the pipeline to Arabic and dialectal misinformation.

Data and code availability

LIAR and FakeNewsNet are publicly available from their authors. The complete experimental code (`cawe_x_experiments.py`), including data download, cleaning, leakage control, modelling, evaluation and figure generation, is provided as supplementary material.

References

1. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1), Article 30.
2. Alghamdi, J., Lin, Y., & Luo, S. (2022). A comparative study of machine learning and deep learning techniques for fake news detection. *Information*, 13(12), Article 576.
3. Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. *Knowledge-Based Systems*, 274, Article 110642.
4. Alghamdi, J., Luo, S., & Lin, Y. (2024). A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83(17), 51009–51067.
5. Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, Article 101805.
6. Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, 45(3), 354–368.
7. Dwivedi, R., Dave, D., Naik, H., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), Article 194.
8. Ecker, U. K. H., Lewandowsky, S., Cook, J., et al. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29.
9. Feng, L., Ahmed, M. O., Hajimirsadeghi, H., & Abdi, A. (2023). Towards better selective classification. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*.
10. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115, Article 105151.
11. Gawlikowski, J., Tassi, C. R. N., Ali, M., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl. 1), 1513–1589.
12. Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., & Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. In *Proceedings of NAACL-HLT* (pp. 6577–6595).

13. Gruber, S. G., & Buettner, F. (2022). Better uncertainty calibration via proper scores for classification and beyond. In *Advances in Neural Information Processing Systems* 35.
14. Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206.
15. Hagen, M., Fröbe, M., Jurk, A., & Potthast, M. (2022). Clickbait spoiling via question answering and passage retrieval. In *Proceedings of ACL* (pp. 7025–7036).
16. Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, 9(10), Article e20382.
17. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., & Davis, J. (2024). Machine learning with a reject option: A survey. *Machine Learning*, 113(5), 3073–3110.
18. Hu, B., Sheng, Q., Cao, J., et al. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of AAAI*, 38(20), 22105–22113.
19. Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, 3, 111–132.
20. Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2), 657–723.
21. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10, 99129–99149.
22. Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University – Computer and Information Sciences*, 35(2), 757–774.
23. Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of COLING* (pp. 4593–4603).
24. Nauta, M., Trienes, J., Pathak, S., et al. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s), Article 295.
25. Pelrine, K., Imouza, A., Thibault, C., et al. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of EMNLP* (pp. 6399–6429).
26. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14, Article 6086.
27. Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine learning with PyTorch and Scikit-Learn*. Packt Publishing.
28. Raza, S., & Ding, C. (2022). Fake news detection based on news content and social contexts: A transformer-based approach. *International Journal of Data Science and Analytics*, 13(4), 335–362.
29. Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., et al. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1), Article 2400304.
30. Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9), 3211–3260.
31. Traub, J., Bungert, T. J., Lüth, C. T., et al. (2024). Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems* 37.
32. Truică, C.-O., & Apostol, E.-S. (2023). It's all in the embedding! Fake news detection using document embeddings. *Mathematics*, 11(3), Article 508.
33. Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural language processing with Transformers: Building language applications with Hugging Face* (Rev. ed.). O'Reilly Media.
34. Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). deep-significance: Easy and meaningful statistical significance testing in the age of neural networks. *arXiv:2204.06815*.
35. Wang, B., Ma, J., Lin, H., et al. (2024). Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference (WWW)* (pp. 2452–2463).
36. Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *arXiv:2308.01222*.
37. Warner, B., Chaffin, A., Clavié, B., et al. (2025). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of ACL* (Vol. 1).
38. Wu, J., Guo, J., & Hooi, B. (2024). Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks. In *Proceedings of ACM SIGKDD* (pp. 3367–3378).
39. Xu, C., & McAuley, J. (2023). A survey on model compression and acceleration for pretrained language models. In *Proceedings of AAAI*, 37(9), 10566–10575.
40. Yuan, L., Chen, Y., Cui, G., et al. (2023). Revisiting out-of-distribution robustness in NLP: Benchmarks, analysis, and LLMs evaluations. In *Advances in Neural Information Processing Systems* 36 (Datasets and Benchmarks Track).