applied sciences

MDPI

Article

Decision-Threshold Misalignment in Cross-Domain Fake News Detection: An Evaluation of Transformer Encoders

Ilkay Yelmen

Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istinye University, Istanbul 34396, Türkiye; ilkay.yelmen@istinye.edu.tr

Abstract

Transformer-based fake news classifiers often perform well within their training domain but show substantial performance degradation when applied to a different domain. This degradation is commonly attributed to changes in learned representations, while the role of the classification decision rule, which assigns the class with the highest predicted probability, receives less attention. In this study, four transformer encoders were trained under a common configuration across four training conditions and five random seeds, yielding 80 training runs. The models were evaluated within their training domains, across domains, and on an external dataset. During preprocessing, source-identifying terms such as publisher names, news agency names, and website domains were removed from the texts before duplicate records were eliminated within and across datasets. Prior to cleaning, a classifier relying on source-identifying terms achieved a Macro-F1 score of 0.994 on ISOT and 0.756 on the pooled dataset. After removing these terms, the pooled score decreased to 0.355, reflecting a class-prior baseline rather than meaningful classification performance. In the most severely affected transfer condition, Macro-F1 decreased to 0.217, although discrimination remained above chance. Calibration error increased by a factor of 2.2 to 4.4. The models classified 82–95% of the target examples as fake, although only 20% were actually fake. Post-hoc class-weight compensation improved performance in the most severely affected transfer direction but had a near-zero average effect overall. Additional cross-fitted threshold analysis recovered approximately 63% of the observed Macro-F1 loss in PolitiFact → GossipCop, compared with approximately 1% in the reverse direction. Class-conditional score and representation analyses showed that the observed changes could not be explained by class-prior differences alone. Overall, the findings indicate that decision-rule misalignment contributes to cross-domain performance degradation, but its effect varies substantially with transfer direction.

Keywords: fake news detection; misinformation detection; cross-domain generalization; domain adaptation; transformer models; model calibration; label shift; data leakage; text classification

Check for updates

Academic Editor: Dong Hyun Jeong

Received: 24 August 2026
Revised: 28 October 2026
Accepted: 2 November 2026
Published: 7 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The reliability of online information is increasingly challenged by misinformation. Automated misinformation detection systems play an important role in identifying potentially misleading content. Research in this field has revealed a wide range of methods that utilize various information sources, such as news content, social context, and dissemination patterns [1]. Text-based Transformer classifiers are widely used because they do not require social context data and can be deployed independently of social media platforms.

The performance of these classifiers in domain-specific evaluations is often very good, but declines sharply when applied to another domain. This decline in performance is well-known and often attributed to domain-specific learned representations that do not generalize well to other domains [2,3]. As a consequence, most proposed solutions focus on adapting the architecture or the representations to account for domain differences.

Testing this hypothesis, however, presents two methodological challenges. First, as noted earlier, limitations in current evaluation protocols make it difficult to determine whether reported performance is indicative of the understanding of content. Previous work has found that fake news classification performance is sensitive to the particular dataset, the evaluation design, and the metric used [4]. Second, although it is well documented in the general machine learning literature that classifier probabilities are distorted under distribution shifts [5,6], the contribution of this phenomenon to cross-domain failure in fake news detection has not been disentangled. The positioning of the decision threshold relative to the source distribution can lead to systematic error if the class prior differs in the target domain, and this error is not apparent in metrics independent of the threshold.

This study assesses cross-domain fake news detection in a leakage-controlled protocol, covering both transfer directions, and analyzing discriminative degradation, calibration drift, and decision-rule misalignment, in addition to representation-level differences. The research question is formulated as follows: To what extent do decision-rule and calibration misalignment contribute to the Macro-F1 degradation observed under domain shift, in addition to the changes in threshold-independent discriminative performance?

This study tests the hypothesis that loss of Macro-F1 across domains is not only due to loss of discriminative performance, but also includes a component that is threshold-sensitive and whose magnitude varies with the direction of transfer. This hypothesis is tested by comparing the performance of threshold-dependent and threshold-independent metrics, measuring the performance that can be recovered by cross-fitted threshold adjustment, and studying the changes in calibration and the distribution of predictions across the different transfer directions.

The contributions of this study are as follows:

- The classification performance provided solely by source-based shortcut terms was measured and compared before and after cleaning; this measurement provides direct evidence of the extent to which the dataset's performance relies on non-content-related cues.
- Duplicate records within and across datasets were removed before partitions were created; it was verified that there was zero overlap between the external validation set and the training set, thereby confirming the independence of the external evaluation set from the training data.
- The transfer was reported separately for both directions; it was demonstrated that the model ranking can be reversed depending on the direction, and it was shown that one-way evaluations can be misleading in model selection.
- The threshold-independent discrimination power, decision-based performance, calibration error, and the performance effect of post-hoc class-weight compensation were jointly evaluated to assess the contribution of calibration and decision-rule misalignment to the observed performance loss.
- No hyperparameter tuning was performed per model, and this constraint is enforced at the code level through automatic checks that trigger a failure if it is violated.

The rest of the paper is organized as follows: Section 2 reviews the relevant literature, presents the benchmark datasets used in the experiments, and describes the research gap that is filled by this paper; Section 3 describes the experimental design, the models used in the experiments, and the statistical analysis used; Section 4 presents the results of the

experiments; Section 5 discusses the results, in the context of previous work, and describes the limitations of this study; Section 6 summarizes the results of this study and outlines future work.

2. Literature Review

2.1. Text-Based Fake News Detection

Most existing fake news detection methods can be categorized into three different approaches: those using only news content, those using social context and information on how the news was disseminated, and multimodal approaches using a combination of both types of information. Social context information contains useful unique information but can be very difficult to use in practice, as it may not exist at the time of publication of the news and requires access to social media platform data. For these reasons, classifiers using only text content are preferred for early detection as well as for use in platforms where social media information is not available.

Text-based fake news detection has been investigated across multiple languages and platforms. A study of Spanish-language misinformation detection evaluated BERT and RoBERTa-based models on political news, demonstrating their applicability to this linguistic setting [7]. For Romanian, both direct fine-tuning and back-translation-based data augmentation have been tested [8,9]; deep learning architectures have been evaluated on Arabic tweet data [10]. A cloud-based system using BERT has also been proposed to analyze sentiment associated with fake news on X (Twitter) [11]. Recent research has also explored XLNet-based architectures incorporating attention and contextual embeddings for misinformation detection [12]. A common feature of these studies is that the evaluation is largely limited to a single dataset and domain-specific tuning.

2.2. Pre-Trained Transformer Encoders

Pre-trained encoders have become the de facto standard in text classification. BERT has popularized the learning of contextual representations through bidirectional masked language modeling [13]. RoBERTa demonstrated that changes to pretraining data, batch size, training duration, and masking strategy can substantially improve BERT-based representations [14]. ELECTRA improves pretraining efficiency through a replaced-token detection objective [15]. Distillation-based approaches reduce the number of parameters and inference cost while preserving a significant portion of the teacher model's performance [16]. DeBERTa proposed an attention mechanism that separates content and position representations [17]; more recent encoders, meanwhile, combine efficiency gains with support for long context [18].

These encoders are used in a wide range of text classification tasks beyond fake news detection; it has been shown that deep learning architectures combined with lexical ontologies improve performance in multi-class document classification [19]. Adaptations of the attention mechanism specifically tailored for fake news detection have also been proposed [20].

The fine-tuning behavior of architectures has been the subject of a separate study. It has been shown that in long-text classification, the pruning strategy, learning rate, and layer-based adjustments significantly affect performance [21]. This observation implies that comparing different architectures under a common hyperparameter budget measures not only architectural differences but also hyperparameter sensitivity. Model-independent explainability methods are used to examine which input components a model relies on to make its decisions; Shapley additive explanations (SHAP) are widely applied for this purpose in data types other than text, such as clinical decision support models [22].

The attention-based architectures can be applied to visual information processing. Recently, a multimodal framework for AI-generated video detection was proposed in [23], which uses spatial, frequency-domain and optical-flow features and applies cross-attention to detect inconsistencies between appearance and motion.

Another study proposed Mutual Retinex, which integrates global feature modeling using Transformer with local feature extraction using CNNs for low-light and underwater image enhancement [24]. These studies are for different tasks but illustrate the use of attention mechanisms and complementary feature representations for different data modalities.

2.3. Cross-Domain and Multi-Domain Detection

The degradation of classification performance across domains is a recurring finding in this literature. Studies on cross-domain fake news detection have shown that differences in distribution across domains reduce performance and have proposed solutions that explicitly model these domain differences [2]. In a multi-domain setting, expert-blend architectures have been developed that distinguish between common and domain-specific patterns across domains [3]. A common assumption in these studies is that the performance loss stems from a problem at the representation level and, therefore, the solution must also be sought at the representation level.

It should also be emphasized that the term "cross-domain" in this literature encompasses distinctly different protocols. Some studies access unlabeled records from the target domain and select a subset of them for labeling [2]; in multi-domain models, however, the target domain is already included in the training mix [3]. In zero-shot transfer, a model is trained only on source-domain data. It is then evaluated on an unseen target dataset. Differences in style, topic, and class distribution can substantially reduce performance [25]. This distinction means that cross-domain values reported by different studies cannot be directly compared: the value produced by a setup with access to the target domain is not on the same scale as the value produced by a setup that uses no target domain data.

This assumption has not been directly tested. Performance loss is often reported using a single decision-based metric. Without threshold-independent metrics, discrimination loss cannot be clearly distinguished from decision-rule misalignment. Furthermore, transfer is mostly reported in a single direction or by averaging across directions, which obscures direction-dependent effects.

2.4. Dataset Bias, Leakage, and Benchmark Datasets

In the area of fake news classification, existing studies that have developed evaluation frameworks have highlighted the large variation in the performance of classifiers depending on the dataset, the evaluation setup and the metric used [4]. This variation is driven by two mechanisms that have already been documented in the general machine learning literature.

First, there is content overlap between the training and the test sets of data. When no cross-validation is used, a model can recognize content from the training set in the test set and thus get a high score. It has been shown that such content overlap has a positive effect on the performance of a model in the evaluation of natural language processing tasks [26]. Second, there are superficial features that are strongly correlated with the class label of a data point but do not represent the content of the data. Such shortcut learning is a well-known cause of poor generalization of deep learning models [27]. In the case of datasets that consist of real news from certain agencies and fake news from certain websites, the agency or the publisher of the news can even be used as a single feature to determine the class label of a data point. Such dataset bias is not specific to the fake news domain and it has already been shown that even small changes in correlation and diversity can cause measurable bias problems [28].

While the existence of these risks is conceptually recognized, quantitatively measuring their magnitude across relevant datasets is not a common practice. This means that comparing performance metrics reported by different studies is risky, because of possible differences in cleaning protocols.

The benchmark datasets used in the study are the following: FakeNewsNet [29], which contains two natural domains, PolitiFact (politics) and GossipCop (gossip); LIAR [30], a dataset of short claims, all from PolitiFact, used as a robustness metric as it represents a different text regime than the long news articles in the dataset; and ISOT [31], a dataset of news articles, independent of the training data used in the study, and thus used as an external validation set.

2.5. Calibration, Prior Shift, and Decision-Rule Misalignment

It has been shown that the prediction probabilities of modern neural networks tend to be overly confident, and that accuracy and calibration can diverge [5]. This tendency becomes more pronounced under distribution shift: a systematic evaluation conducted across numerous model families and types of shift revealed that the reliability of predicted uncertainty decreases as the shift increases [6]. Recent studies propose methods that address out-of-distribution detection and confidence calibration together [32], examine the identification of out-of-distribution examples in text data using statistical techniques [33], and investigate the interaction between class imbalance and out-of-distribution detection [34].

In label shift—a specific form of drift—class-conditional input distributions are preserved while class prior distributions change. In this case, it has long been known that the outputs of a classifier trained on the source distribution must be corrected according to the new prior, and that this correction can be performed without retraining the classifier [35]. Domain adaptation under label shift is also the subject of current studies; approaches based on conditional support alignment report performance improvements in this setting [36]. The expected calibration error is widely used to assess predictive calibration [5]. Bayesian binning has also been used as a calibration method for predicted probabilities [37]. To compare multiple conditions and models, non-parametric rank-based tests and multiple comparison corrections are recommended [38]. Differences in class prevalence alone do not establish label shift, as this requires approximately invariant class-conditional input distributions. Hence, we distinguish observed class-prior mismatch from verified label shift and evaluate the latter assumption through class-conditional distribution diagnostics.

Fake news detection in different domains and calibration under distribution shift are two fields of research that have so far developed largely in isolation from each other. Calibration metrics are rarely reported in the former field, while calibration research often focuses on image classification and general text tasks.

This paper presents a novel approach for evaluating cross-domain fake news detection models. Unlike previous work that has attempted to improve transfer performance via architectural, representation-level, or domain adaptation methods, this work focuses on validly interpreting the results of cross-domain evaluations. The evaluation framework presented in this work incorporates leakage and shortcut auditing, bidirectional transfer with no access to target-domain data, threshold-independent discrimination, calibration analysis, prediction distribution analysis, and controlled ablations into one comprehensive evaluation framework. The results of this study can be used to determine if poor decision-based performance under domain shift is due solely to a reduction in discriminative power or both a reduction in discriminative power and misalignment of the decision rule and calibration.

The above synthesis points to three gaps. First, quantitative reporting of duplicate and leakage control remains limited in cross-domain fake news evaluation. Second, transfer

is rarely reported separately in both directions. Third, discriminative degradation and calibration/decision-rule misalignment are rarely examined jointly.

This study addresses these three gaps within a single controlled experimental design by quantitatively auditing duplicates and shortcut signals, reporting transfer separately in both directions, and jointly examining threshold-independent discrimination, decision-based performance, and calibration.

3. Materials & Methods

3.1. Dataset

In this study, three different datasets were used: FakeNewsNet, LIAR, and ISOT. FakeNewsNet was used for the primary cross-domain evaluation because it includes two natural news domains: PolitiFact and GossipCop. LIAR was included in the secondary robustness evaluation because it presents a different text structure consisting of short political claims. ISOT, on the other hand, was used for external evaluation because it is an independent data source.

The official FakeNewsNet release does not distribute news body text; therefore, this study used the publicly available MM-SOC derivative datasets [39]. Specifically, PolitiFact was obtained from the Ahren09/MMSoc_PolitiFact Hugging Face repository (commit 6228a621f5e2bedf8aa715bdb0bb35707e204a9f) and GossipCop from Ahren09/MMSoc_GossipCop (commit 67a204e93e6bc058dab3a8bad118218707c923c5). These versioned releases contain 483 PolitiFact records and 12,660 GossipCop records. Accordingly, the results reported in this study apply to these specific MM-SOC subsets rather than to the complete canonical FakeNewsNet distribution. Data inspection identified two records with identical links and headlines but conflicting Fake and Real labels. Since they contained conflicting label information, both copies of these records were removed from the dataset.

The LIAR dataset originally contains six different truth labels. To establish a binary classification framework consistent with the other datasets in this study, these labels were converted into two classes using a fixed mapping determined before the experiments began. Accordingly, the "pants-fire," "false," and "barely-true" labels were assigned to the "Fake" class, while the "half-true," "mostly-true," and "true" labels were assigned to the "Real" class.

The statistics for the datasets obtained after cleaning and duplicate auditing are presented in Table 1. The FakeNewsNet/PolitiFact dataset consists of 479 records, FakeNewsNet/GossipCop of 12,604, LIAR of 12,805, and ISOT of 38,770. The proportion of the "Fake" class in the datasets is 34.0%, 20.0%, 44.1%, and 46.1%, respectively.

Table 1. Dataset statistics after deduplication and leakage cleaning.

Dataset/Domain	Records	Real	Fake	Fake Ratio (%)	Mean Length (Chars)	Median Tokens	Truncated at 128 Tokens (%)
FakeNewsNet/PolitiFact	479	316	163	34.0	6710	564	77.0
FakeNewsNet/GossipCop	12,604	10,084	2520	20.0	3200	498	88.4
LIAR	12,805	7154	5651	44.1	106	23	0.0
ISOT	38,770	20,900	17,870	46.1	2463	—	—

Table 1 also highlights the significant differences in text lengths and token structures across the datasets. While PolitiFact examples contain an average of 6710 characters and a median of 564 tokens, these values for GossipCop examples are 3200 characters and 498 tokens, respectively. LIAR, on the other hand, presents a significantly shorter text regime with an average of 106 characters and a median of 23 tokens. At 128 tokens, 77.0% of

PolitiFact training examples were truncated. The corresponding rate for GossipCop was 88.4%. No truncation was observed in LIAR.

The character counts reported in the table represent the final input to the model, i.e., the combination of the headline and body text. Token counts were calculated using the bert-base-uncased tokenizer, and truncation rates were reported across the training partitions. Since ISOT was used exclusively as an external evaluation set and did not have a training partition within the scope of this study, token counts and truncation rates were not calculated for this dataset. In all evaluations, the "Fake" class was defined as the positive class.

A data split refers to the division of a dataset into training, validation, and test subsets. The PolitiFact and GossipCop datasets were stratified into training (70%), validation (15%), and test (15%) sets while preserving class distribution. For LIAR, the official training, validation, and test sets were preserved; only records removed during duplicate auditing were excluded from the respective sets. All data partitions were generated once independently of random seeds and versioned with content summaries. Thus, five different random seeds were used solely to illustrate the stochastic variability of model training; no changes were made to the data partitions. The final training, validation, and test set sizes are presented in Table 2.

Table 2. Training, validation, and test split sizes for datasets used in model training. Splits are seed-independent and versioned by a content hash.

Split	Train	Validation	Test	Total
PolitiFact	335	71	73	479
GossipCop	8823	1890	1891	12,604
FakeNewsNet source union	9158	1961	1964	13,083
LIAR (official splits)	10,241	1282	1282	12,805

3.2. Data Preprocessing and Leakage Control

The reliability of the evaluation results may be affected by the model's ability to recognize content it encountered during training or by its reliance on source clues associated with the class label but unrelated to the content. For this reason, an identical preprocessing and validation pipeline, independent of the model, was applied to all datasets.

During the preprocessing stage, empty records were removed, HTML tags were stripped, and elements that could carry source information—such as links, email addresses, and usernames—were replaced with placeholders. Unicode NFKC normalization and whitespace simplification were then applied. Lowercasing, punctuation removal, stop-word filtering, stemming, and lexicalization were not performed. The reason for this is that the Transformer models used employ subword-based tokenizers, and case and punctuation information can serve as meaningful signals for some models. The preprocessing function was designed independently of the model identifier, ensuring identical data transformations across all models.

For FakeNewsNet, the original MM-SOC text field was used directly, with the headline appearing at the beginning of the article. For ISOT, the headline and body were concatenated using a period and a space, whereas LIAR used the statement text alone. Inputs exceeding 128 tokens were truncated from the end, preserving the beginning of each document. The 128-token limit included two special tokens, leaving up to 126 content tokens. No model-specific delimiter tokens were inserted.

After source-identifying terms had been removed from the texts, duplicate records were identified and removed within and across datasets before the data splits were created. Exact duplicates were identified using the SHA-256 hashes of the normalized texts. Near-duplicate candidates were identified using MinHash with locality-sensitive hashing

(MinHash-LSH), a method for efficiently identifying similar documents. Candidate pairs were then represented using character-level 5-grams weighted by term frequency–inverse document frequency (TF-IDF). Cosine similarity was used to measure their textual similarity. A threshold of 0.85 was applied for Jaccard similarity, and the identified near-duplicates were removed from the dataset.

To control for source-based shortcut signals, a 53-term dictionary was created consisting of publisher names, link domain names, verification site names, and news agency bylines. These terms were removed from the model inputs. To determine the extent to which these terms alone contributed to classification performance, a logistic regression classifier was trained on a bag-of-words representation consisting solely of these terms, and performance was measured using Macro-F1 with three-fold cross-validation. This classifier is not part of the main experimental models; it was used solely to quantitatively assess data quality and the magnitude of source-based shortcut signals.

The results of the reference classifier based on shortcut terms are presented in Table 3 and Figure 1. A Macro-F1 score of 0.756 was achieved on the pooled data prior to cleaning. Marked differences were observed across the datasets: Macro-F1 was 0.994 for ISOT, 0.629 for GossipCop, 0.515 for PolitiFact, and 0.361 for LIAR. The proportion of records containing shortcut terms also showed a similar distribution. While 77.8% of ISOT records contained at least one shortcut term, this proportion was 28.8% for GossipCop, 27.7% for PolitiFact, and only 0.4% for LIAR. The record counts in Table 3 reflect the values prior to duplicate removal, as the shortcut check was performed on the entire dataset before duplicate removal.

Table 3. Shortcut-only baseline using 53 source-related terms with 3-fold cross-validation.

Scope	Records	Records Containing Shortcut Terms	Share of Records (%)	Shortcut-Only Macro-F1
Pooled (all datasets)	70,877	38,744	54.7	0.756
ISOT	44,898	34,912	77.8	0.994
FakeNewsNet/GossipCop	12,660	3651	28.8	0.629
FakeNewsNet/PolitiFact	483	134	27.7	0.515
LIAR	12,836	47	0.4	0.361

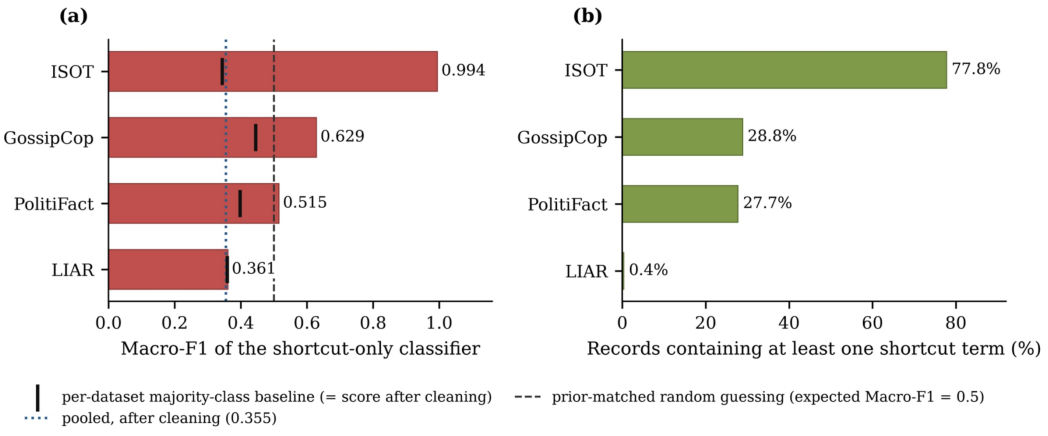


Figure 1. Shortcut audit. (**a**) Macro-F1 of a classifier that sees only the 53 source-related terms, per dataset, against the pooled score obtained after those terms are removed. (**b**) Proportion of records in which at least one shortcut term occurs. ISOT separates sharply from the remaining corpora on both panels.

After the shortcut terms were removed, the shortcut feature matrix contained no non-zero features. The resulting pooled Macro-F1 of 0.355 therefore corresponds to the degenerate class-prior baseline rather than to classification based on shortcut features.

During duplicate auditing, 64,658 of 70,877 records were retained; 5850 records were removed as exact duplicates and 369 as near-duplicates. Within ISOT, 6128 pairwise exact-overlap matches were identified; this count refers to duplicate matches rather than to the number of removed records. No overlap was found between FakeNewsNet and ISOT, confirming the independence of the external evaluation set from the training data.

3.3. Training and Experimental Setup

In this study, four Transformer encoders were evaluated under four different training conditions with the same training budget. Each condition was repeated with five independent random seeds, resulting in a total of 80 training runs. The general workflow of the experimental process is shown in Figure 2.

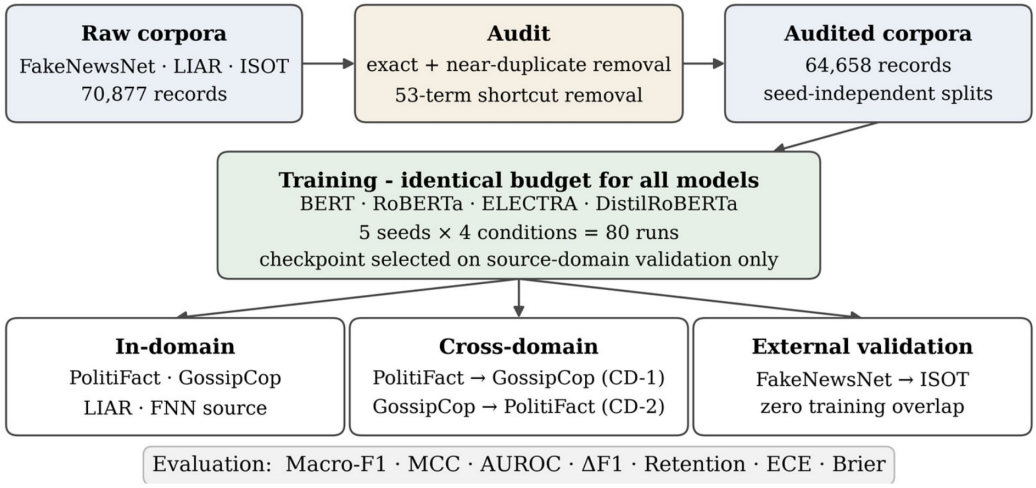


Figure 2. Experimental pipeline for leakage-controlled in-domain, cross-domain, and external evaluation.

The experimental conditions are presented in Table 4. No separate model training was performed for cross-domain evaluations. Model checkpoints trained on the source domain and selected solely based on the source domain validation set were evaluated directly on the target domain test set. Thus, no labeled or unlabeled data from the target domain was used in model training, hyperparameter selection, or checkpoint selection. This design ensures that cross-domain results are measured independently of target domain knowledge.

Table 4. Experimental conditions for in-domain, cross-domain, and external evaluation.

Condition	Training Set	Validation Set	Evaluation Sets	Runs
FAM_A	PolitiFact train	PolitiFact val	PolitiFact test (in-domain); GossipCop test (CD-1)	20
FAM_B	GossipCop train	GossipCop val	GossipCop test (in-domain); PolitiFact test (CD-2)	20
FAM_C	LIAR train	LIAR val	LIAR test (in-domain)	20
FAM_D	PolitiFact + GossipCop train	PolitiFact + GossipCop val	Source test (in-domain); ISOT (external)	20

Under these experimental conditions, four Transformer encoders were evaluated: bert-base-uncased, roberta-base, google/electra-base-discriminator, and distilroberta-base.

In the original experimental design, microsoft/deberta-v3-base was planned as the fourth model. However, under the fixed learning rate of 2×10^{-5}, its validation Macro-F1 remained at the majority-class baseline (0.398), while training loss increased across epochs. Similar majority-class predictions were observed with microsoft/deberta-v3-small and in a preliminary float32 run. No skipped optimization steps were observed. Layer-wise learning rate decay was not tested, because model-specific optimization adjustments were

excluded from the predefined common-budget protocol. An additional diagnostic run with a maximum of 10 epochs and patience of 10, using the same learning rate and seed 42, also failed to converge; training loss became NaN in epoch 3. This single-seed diagnostic does not establish an inherent limitation of DeBERTa-v3. The DeBERTa family was therefore excluded from the main comparison and replaced with distilroberta-base.

All models were trained using the identical training configuration provided in Table 5. No hyperparameter tuning was performed per model; this constraint was enforced through automated checks that verified the absence of hyperparameters in the model configuration files and generated errors when attempts were made to override them on a per-model basis.

Table 5. Training configuration, identical for all four models; no per-model tuning was performed.

Parameter	Value
Task head	Sequence classification, num_labels = 2
Maximum sequence length	128 tokens
Effective batch size	16
Optimizer	AdamW (fused CUDA kernel)
Learning rate	2×10^{-5}
Weight decay	0.01
Scheduler	Linear decay with 10% warmup
Gradient clipping	1.0 (global norm)
Maximum epochs	5
Early stopping	Patience 2, monitored on validation Macro-F1
Checkpoint selection	Highest validation Macro-F1
Loss	Class-weighted cross-entropy (weights from training split only)
Mixed precision	bfloat16
Random seeds	42, 52, 62, 72, 82
Inference batch size	128
Decision rule	Argmax (no threshold search)

Class weights were calculated solely from the class frequencies of the relevant training partition; validation and test distributions were not included in this calculation.

The two ablation conditions presented in Section 4.4 were written to separate output roots outside the main matrix and did not affect the main 80 runs in any way. In the first ablation (A1), class weighting was completely removed, and unweighted cross-entropy was used; all other hyperparameters and partitions remained unchanged. In the second ablation (A2), the GossipCop training set was downsampled to the same size (335 examples) as the PolitiFact training set; only the amount of data was changed while preserving GossipCop's own class ratio, and the validation and test sets were left untouched. Both ablations were run with four models and five seeds—that is, 20 runs, just as in the main matrix.

Input-length sensitivity was examined separately from the main 80-run comparison. DistilRoBERTa was trained with maximum sequence lengths of 256 and 512 tokens on both source domains using three random seeds (42, 52, and 62), yielding 12 additional training runs. The corresponding 128-token results were obtained from the original experiments. A complementary inference-time chunking analysis was conducted using all four encoders with seed-42 checkpoints retrained under the original 128-token configuration because the original checkpoints had not been retained. Without further model training, document-level predictions were generated by averaging logits across the first 1, 2, 4, or all available chunks of 126 content tokens. Results were also examined separately for documents identified as not requiring truncation according to the BERT tokenizer. Additional methodological details and results are reported in Supplementary Tables S13 and S14.

During training, steps where the overall gradient norm before clipping was infinite were skipped. This safeguard is necessary: the clipping function applies the total norm as a multiplier to all gradients; if a single corrupted gradient norm becomes infinite or undefined, all parameters are corrupted in a single step. No skipped steps were observed in the reported runs.

Experiments were conducted on a single NVIDIA GeForce RTX 4050 Laptop GPU (6 GB VRAM). Software environment: Python 3.13, PyTorch 2.13.0 (CUDA 12.6), Transformers 5.15.0. For each run, a configuration snapshot, seed, split ID, version information, epoch-based log, metric values, and sample-level predictions (ID, true label, prediction, probability, and logit vectors) were written to disk. A total of 80 runs were completed, with no failed runs observed.

Code validation combined automated protocol checks with independent verification of the saved experimental results. All 11 protocol checks passed. Dataset integrity was examined using record-level hashes, class distributions, and overlap checks. Independent recalculation from saved predictions reproduced the reported performance and calibration results and confirmed consistency across all 80 training runs. Following code refactoring, the analyses were rerun and numerical differences remained below 10^{-14}. Additional calibration, chunked-inference, and representation analyses were checked against existing implementations or saved predictions. Package imports, execution paths, and static-analysis checks were also verified. Together, these procedures provide evidence of implementation consistency and support the reliability of the reported results.

3.4. Evaluation Metrics and Statistical Analysis

This section explains the metrics used to evaluate model performance and the statistical analysis methods applied in comparing models.

Macro-F1 was used as the primary evaluation metric. This metric was chosen because it assigns equal weight to both classes under class imbalance. The F1 score is the harmonic mean of precision and recall. Macro-F1 averages class-specific F1 scores equally, whereas Weighted-F1 weights them by class support, giving greater influence to the majority class. To ensure comparability with the literature, accuracy, precision, and recall values were also reported. However, since accuracy can produce misleading results in favor of the majority class in imbalanced datasets, the Matthews correlation coefficient (MCC) was also reported as a complementary threshold-dependent measure. In addition, two threshold-independent metrics were included: the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC). Precision and recall values were calculated for the positive class, and the "Fake" class was fixed as the positive class in all experiments. Balanced accuracy and weighted F1 were also calculated but were not reported separately in the main tables because they largely overlap with Macro-F1 and accuracy.

Two derived metrics were defined to evaluate transfer performance. Let $F1_{ID}$ denote a model's Macro-F1 score in the source domain and $F1_{OOD}$ its score in the target domain. The performance drop and retained performance ratio were calculated using Equations (1) and (2), respectively.

$$\Delta F1 = F1_{ID} - F1_{OOD} \tag{1}$$

$$\text{Retention} = \frac{F1_{OOD}}{F1_{ID}} \times 100 \tag{2}$$

These metrics were calculated using seed-matched data and then averaged across five seeds.

For calibration, the expected calibration error (ECE) was calculated using 15 equal-width bins, and the Brier score was computed. The calibration analysis is not part of the main performance claim and is reported separately.

To quantify threshold-sensitive performance loss, the original argmax Macro-F1 was compared with the best achievable threshold-based Macro-F1 using saved prediction scores. The observed loss was decomposed into source-threshold, class-prior, residual discrimination-related, and target-threshold components, as detailed in Supplementary Table S1. Because oracle thresholds selected on the evaluation set may be optimistic, stratified two-fold cross-fitting with 20 repetitions was additionally performed, with thresholds selected on one half and evaluated on the other. Balanced accuracy was used as a complementary measure, and bootstrap confidence intervals were calculated for the oracle estimates. These diagnostic comparisons do not represent target-free adaptation or uniquely identifiable causal contributions. Post-hoc class-weight compensation was used solely as an inference-time diagnostic intervention. Its purpose was to examine the sensitivity of classification decisions to source-derived class weights without retraining the models. It was not intended as a general-purpose performance improvement or a validated correction for target-domain class priors.

Uncertainty in the threshold-sensitive performance estimates was quantified using hierarchical bootstrap resampling across training seeds and test instances. A total of 1000 bootstrap repetitions were performed, except for the ISOT evaluation, which used 200 repetitions. During cross-fitting, repeated copies of the same test instance were assigned to the same partition to prevent overlap between threshold selection and evaluation. Exploratory one-sided Wilcoxon signed-rank tests were applied to run-level estimates, with Holm correction across 20 mechanism-related tests. Calibration sensitivity was assessed using equal-width ECE with 10, 20, and 30 bins, equal-mass and debiased estimators, kernel-based calibration error, Brier score, and negative log-likelihood. Detailed procedures and results are provided in the Supplementary Materials.

To distinguish class-prior mismatch from pure label shift, class-conditional prediction-score distributions were compared using Kolmogorov–Smirnov statistics, Wasserstein distance, and Jensen–Shannon divergence. Within-class domain discrimination was additionally assessed using TF-IDF features and Transformer representations, together with maximum mean discrepancy tests. Synthetic label-shift controls and exploratory feature–label association analyses were used to examine whether the observed differences could be explained by changes in class prevalence alone. Detailed procedures and results are provided in the Supplementary Materials.

Overall differences between models were tested using the Friedman test, in which blocks (scenario × seed) and treatments (model) were defined. When the test was significant, paired Wilcoxon signed-rank tests were applied, and the Holm correction was used for multiple comparisons. Effect size was reported using the matched-pairs rank-biserial correlation. The significance level was set at $\alpha = 0.05$. If any cells were missing, the test was not performed and an error was generated; missing values were not imputed with the mean.

4. Experimental Results

4.1. Model Performance Results

This section presents the in-domain, cross-domain, and external evaluation results. Model performance is compared across these conditions.

The in-domain performance results are presented in Table 6. The four Transformer models generally performed similarly across all datasets. On PolitiFact, the difference between the highest and lowest average Macro-F1 values is 0.026, while on GossipCop it is only 0.012. The fact that these differences fall within the range of their respective standard deviations indicates that there is no clear superiority among the models under in-domain conditions.

Table 6. In-domain performance over five random seeds (mean ± standard deviation).

Dataset	Model	Accuracy	Precision (Fake)	Recall (Fake)	Macro-F1	MCC	AUROC	AUPRC
PolitiFact	BERT	0.803 ± 0.021	0.753 ± 0.032	0.640 ± 0.120	0.770 ± 0.038	0.552 ± 0.056	0.842 ± 0.053	0.767 ± 0.065
	RoBERTa	0.800 ± 0.039	0.694 ± 0.067	0.752 ± 0.059	0.782 ± 0.041	0.568 ± 0.083	0.898 ± 0.033	0.788 ± 0.056
	ELECTRA	0.789 ± 0.037	0.731 ± 0.078	0.616 ± 0.061	0.756 ± 0.041	0.520 ± 0.084	0.887 ± 0.020	0.796 ± 0.033
	DistilRoBERTa	0.803 ± 0.028	0.728 ± 0.063	0.688 ± 0.066	0.778 ± 0.030	0.560 ± 0.061	0.870 ± 0.019	0.755 ± 0.034
GossipCop	BERT	0.850 ± 0.005	0.617 ± 0.019	0.662 ± 0.036	0.772 ± 0.005	0.545 ± 0.011	0.868 ± 0.012	0.707 ± 0.020
	RoBERTa	0.862 ± 0.022	0.669 ± 0.085	0.652 ± 0.061	0.784 ± 0.018	0.573 ± 0.033	0.879 ± 0.006	0.733 ± 0.018
	ELECTRA	0.863 ± 0.009	0.672 ± 0.040	0.623 ± 0.036	0.780 ± 0.007	0.562 ± 0.015	0.864 ± 0.004	0.714 ± 0.012
	DistilRoBERTa	0.859 ± 0.012	0.654 ± 0.046	0.648 ± 0.059	0.780 ± 0.011	0.562 ± 0.020	0.874 ± 0.008	0.722 ± 0.008
LIAR	BERT	0.637 ± 0.010	0.586 ± 0.010	0.556 ± 0.068	0.627 ± 0.014	0.257 ± 0.028	0.672 ± 0.012	0.588 ± 0.011
	RoBERTa	0.643 ± 0.012	0.600 ± 0.022	0.536 ± 0.024	0.631 ± 0.008	0.265 ± 0.019	0.684 ± 0.010	0.613 ± 0.012
	ELECTRA	0.642 ± 0.011	0.607 ± 0.020	0.501 ± 0.055	0.626 ± 0.015	0.261 ± 0.025	0.684 ± 0.018	0.605 ± 0.016
	DistilRoBERTa	0.646 ± 0.011	0.607 ± 0.016	0.524 ± 0.023	0.633 ± 0.011	0.270 ± 0.022	0.687 ± 0.014	0.606 ± 0.020
FakeNewsNet source	BERT	0.843 ± 0.010	0.625 ± 0.046	0.607 ± 0.053	0.757 ± 0.004	0.517 ± 0.008	0.853 ± 0.010	0.681 ± 0.018
	RoBERTa	0.864 ± 0.006	0.698 ± 0.037	0.605 ± 0.042	0.781 ± 0.003	0.567 ± 0.007	0.875 ± 0.004	0.732 ± 0.009
	ELECTRA	0.858 ± 0.011	0.675 ± 0.048	0.607 ± 0.052	0.774 ± 0.006	0.552 ± 0.012	0.864 ± 0.005	0.714 ± 0.006
	DistilRoBERTa	0.859 ± 0.006	0.669 ± 0.029	0.623 ± 0.045	0.778 ± 0.007	0.558 ± 0.013	0.874 ± 0.004	0.724 ± 0.009

On the LIAR dataset, Macro-F1 values ranged from 0.626 to 0.633 and were lower than those observed in other in-domain conditions. This result suggests that LIAR's structure, consisting of short political claims, presents a different classification problem compared to long news texts.

The effect of class imbalance is particularly evident in the GossipCop results. Although accuracy in this dataset ranges from 0.850 to 0.863, Macro-F1 is only between 0.772 and 0.784. Similarly, for the "Fake" class—defined as the positive class—precision ranges from 0.617 to 0.672, while recall ranges from 0.623 to 0.662. This discrepancy between accuracy and Macro-F1 highlights that evaluating model performance based solely on the accuracy metric in cases of class imbalance may overstate the model's actual performance.

The transfer results are presented in Table 7, the degradation based on in-domain performance in Table 8, and a comparative view with in-domain performance in Figure 3. Related analyses of threshold-sensitive performance loss and statistical uncertainty are provided in Supplementary Tables S1, S2 and S9. A substantial loss in performance is observed in both directions, but the magnitude of the loss varies sharply depending on the direction. Retention ranges from 28.1% to 42.5% in the PolitiFact → GossipCop direction, while it ranges from 60.2% to 77.8% in the GossipCop → PolitiFact direction.

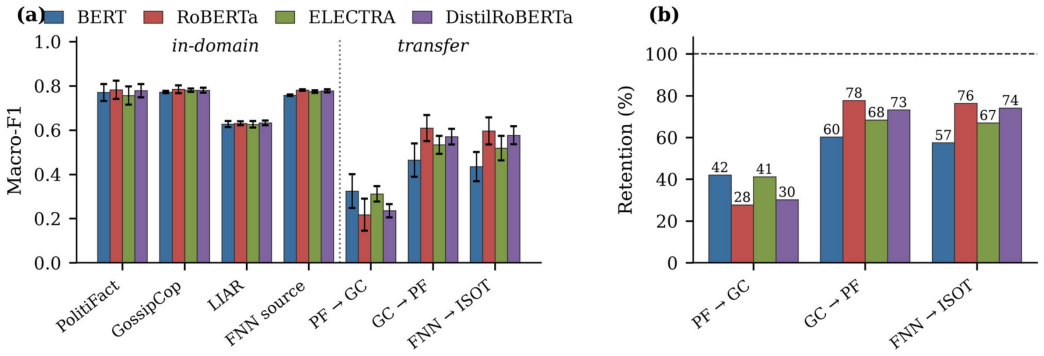


Figure 3. Performance under domain shift: (**a**) Macro-F1 across in-domain and transfer conditions; (**b**) retained performance relative to the corresponding in-domain score. Error bars indicate one standard deviation over five seeds.

The model rankings are reversed between the directions. RoBERTa achieves the highest Macro-F1 (0.609) in the GossipCop → PolitiFact direction, while producing the lowest value (0.217) in the PolitiFact → GossipCop direction. BERT exhibits the exact opposite behavior in these two directions.

BERT produced MCC values close to zero in these settings, indicating weak agreement between predictions and true labels.

Reporting precision and recall separately makes the nature of the loss immediately apparent. In the PolitiFact → GossipCop direction, recall for the "Fake" class rises to 0.877–0.964, while precision remains at 0.204–0.215—a level close to the baseline rate of "Fake" in the target set. This combination is the expected signature of a situation where the models label nearly all examples as "Fake": nearly all actual "Fake" examples are correctly identified, but only one-fifth of the positive predictions turn out to be correct.

The joint evaluation of threshold-dependent and threshold-independent metrics provides further insight into performance degradation under domain shift. In Figure 4, results for the transfer from PolitiFact to GossipCop fall well below the diagonal, indicating substantially lower Macro-F1 values than those observed under in-domain conditions, despite comparable AUROC values. Supplementary Table S7 provides synthetic prior-shift controls, while Table S9 reports discrimination and threshold-sensitivity estimates supporting this interpretation.

Table 7. Cross-domain and external validation performance over five random seeds (mean ± standard deviation).

Transfer	Model	Accuracy	Precision (Fake)	Recall (Fake)	Macro-F1	MCC	AUROC	AUPRC
PolitiFact → GossipCop	BERT	0.333 ± 0.087	0.215 ± 0.007	0.877 ± 0.126	0.324 ± 0.076	0.082 ± 0.011	0.623 ± 0.032	0.282 ± 0.028
	RoBERTa	0.238 ± 0.061	0.204 ± 0.011	0.964 ± 0.026	0.217 ± 0.073	0.014 ± 0.066	0.662 ± 0.024	0.321 ± 0.065
	ELECTRA	0.314 ± 0.032	0.213 ± 0.007	0.903 ± 0.025	0.311 ± 0.035	0.077 ± 0.030	0.626 ± 0.020	0.282 ± 0.019
	DistilRoBERTa	0.249 ± 0.024	0.204 ± 0.003	0.952 ± 0.021	0.235 ± 0.030	0.040 ± 0.017	0.679 ± 0.018	0.360 ± 0.048
GossipCop → PolitiFact	BERT	0.485 ± 0.091	0.328 ± 0.085	0.432 ± 0.033	0.465 ± 0.075	−0.050 ± 0.139	0.491 ± 0.039	0.380 ± 0.034
	RoBERTa	0.660 ± 0.063	0.516 ± 0.089	0.440 ± 0.094	0.609 ± 0.059	0.228 ± 0.120	0.603 ± 0.057	0.506 ± 0.059
	ELECTRA	0.573 ± 0.058	0.397 ± 0.049	0.432 ± 0.095	0.533 ± 0.041	0.079 ± 0.075	0.540 ± 0.058	0.427 ± 0.056
	DistilRoBERTa	0.616 ± 0.026	0.437 ± 0.040	0.432 ± 0.095	0.571 ± 0.035	0.145 ± 0.070	0.532 ± 0.072	0.429 ± 0.079
FakeNewsNet → ISOT	BERT	0.439 ± 0.061	0.414 ± 0.050	0.519 ± 0.087	0.435 ± 0.065	−0.115 ± 0.124	0.442 ± 0.078	0.485 ± 0.065
	RoBERTa	0.601 ± 0.063	0.578 ± 0.090	0.599 ± 0.079	0.596 ± 0.061	0.207 ± 0.118	0.632 ± 0.050	0.671 ± 0.040
	ELECTRA	0.523 ± 0.051	0.490 ± 0.043	0.630 ± 0.062	0.518 ± 0.055	0.061 ± 0.096	0.576 ± 0.047	0.618 ± 0.042
	DistilRoBERTa	0.581 ± 0.036	0.537 ± 0.033	0.700 ± 0.063	0.576 ± 0.040	0.183 ± 0.063	0.665 ± 0.030	0.694 ± 0.035

Table 8. Transfer degradation based on seed-paired Macro-F1 loss (ΔF1) and retained performance (Retention). ↓ indicates that lower values are better, whereas ↑ indicates that higher values are better.

Transfer	Model	ΔF1 ↓	Retention (%) ↑
PolitiFact → GossipCop	BERT	+0.446	42.5
	RoBERTa	+0.565	28.1
	ELECTRA	+0.445	41.2
	DistilRoBERTa	+0.543	30.2
GossipCop → PolitiFact	BERT	+0.307	60.2
	RoBERTa	+0.175	77.8
	ELECTRA	+0.247	68.4
	DistilRoBERTa	+0.209	73.2
FakeNewsNet → ISOT	BERT	+0.322	57.5
	RoBERTa	+0.185	76.3
	ELECTRA	+0.255	67.0
	DistilRoBERTa	+0.201	74.1

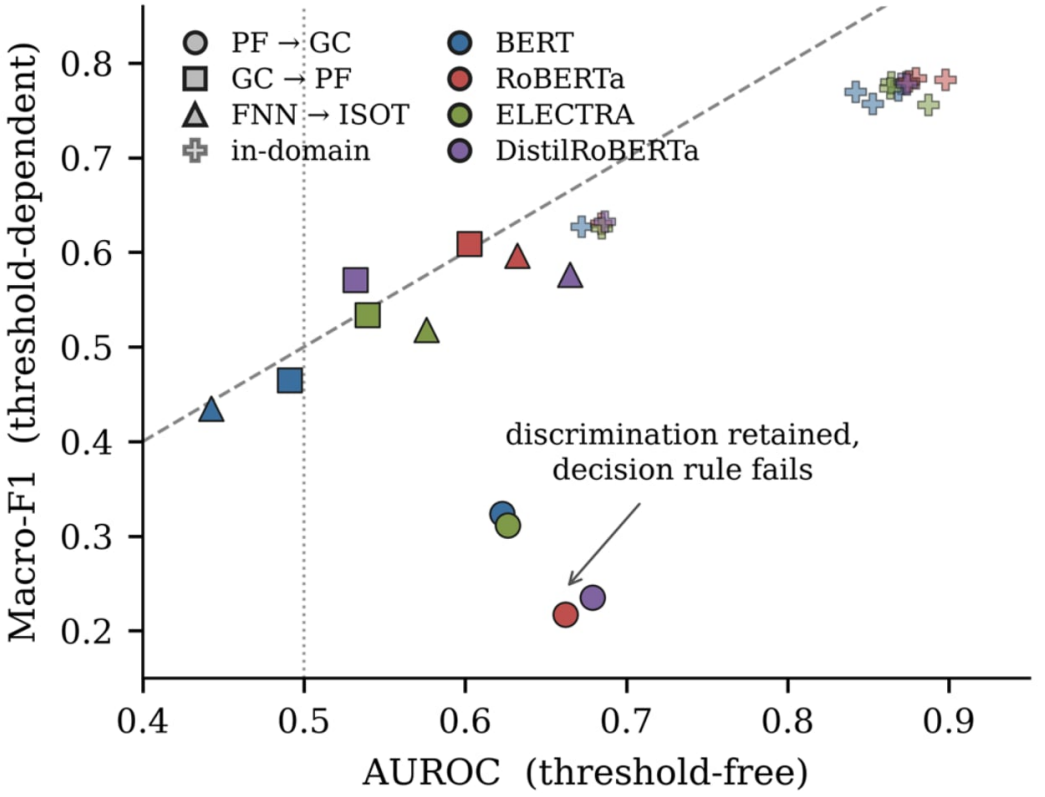


Figure 4. Relationship between AUROC and Macro-F1 across evaluation settings, highlighting decision-rule misalignment under domain shift. Colors indicate models, and marker shapes indicate evaluation settings. The dashed diagonal line represents Macro-F1 = AUROC (visual reference only), while the vertical dotted line marks chance-level AUROC (0.5).

An analysis of the prediction distributions shows that models trained on PolitiFact classify 82–95% of the examples in the GossipCop test set as "Fake"; the actual "Fake" rate in this set is 20%. In contrast, under the same conditions, the AUROC ranges from 0.623 to 0.679 and therefore remains above the 0.5 chance reference.

No such deviation is observed under in-domain conditions: the prediction distributions of models trained on GossipCop and evaluated on GossipCop are close to the true distribution (a "Fake" rate of 0.20 versus 0.19–0.22). The comparison is presented in

Figure 5. Supplementary Tables S7 and S9 provide complementary evidence on predicted class proportions and distribution shifts.

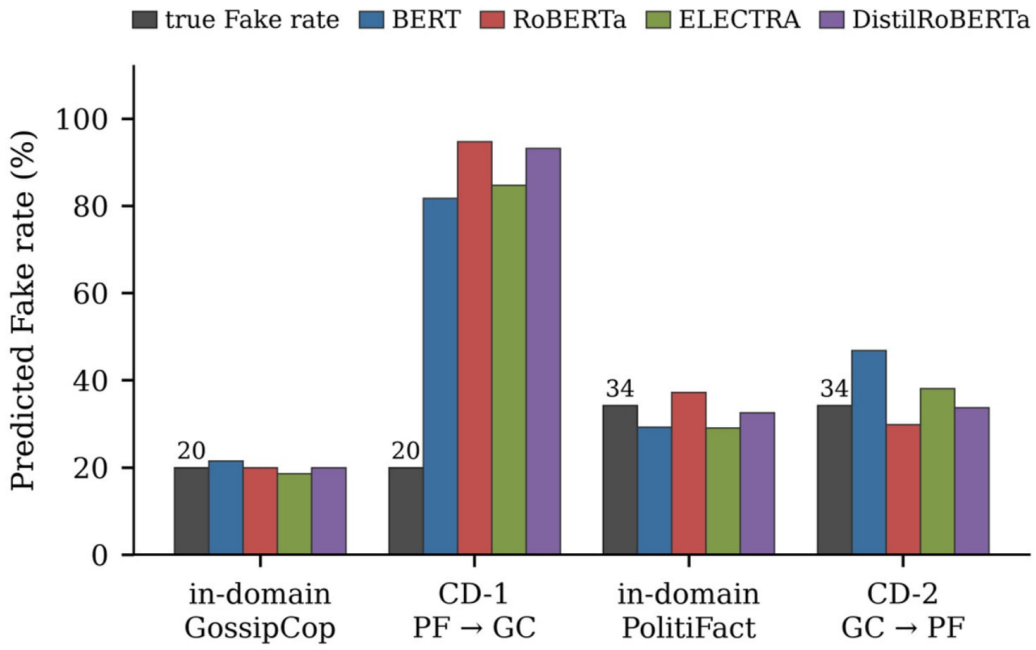


Figure 5. Predicted versus true Fake-class proportions across evaluation settings, showing decision-distribution collapse under CD-1.

4.2. Calibration and Class-Weight Correction Analysis

To further examine the degradation in decision-based metrics, probability calibration and the effect of post-hoc class-weight compensation at inference were analyzed separately.

The calibration results are presented in Table 9. Under in-domain conditions, ECE ranges from 0.044 to 0.215. Under distribution shift, this value increases significantly across all three conditions. Based on the original ECE calculation using 15 equal-width bins, the mean ECE under domain shift was 4.4 times the in-domain mean for BERT, 4.0 times for ELECTRA, 3.8 times for RoBERTa, and 2.2 times for DistilRoBERTa. Sensitivity analyses using alternative binning schemes and calibration estimators, together with Brier score and negative log-likelihood, consistently indicated deterioration under domain shift. However, the estimated magnitude of this deterioration varied across the measures considered.

The highest ECE value is 0.603, produced by RoBERTa in the PolitiFact → GossipCop direction; Macro-F1 also reaches its lowest value (0.217) for the same model and direction. The reliability diagrams are presented in Figure 6. Additional calibration sensitivity analyses complementing the results in Table 9 and the reliability diagrams in Figure 6 are provided in Supplementary Tables S11 and S12.

As a post-hoc diagnostic correction, the predicted class probabilities were divided by the corresponding class weights used during training and subsequently renormalized. This correction uses no information about the target domain and contains no free parameters; it relies solely on the statistics of the training partition of the source domain. It should therefore be interpreted as an inference-time compensation for the decision-boundary shift associated with class-weighted training rather than as an exact reversal of the effects of weighted optimization. The effect of this correction on Macro-F1, expressed as the difference relative to the argmax decision rule, is presented in Table 10 and Figure 7. Supplementary Table S3 further examines how this effect varies with predicted class imbalance.

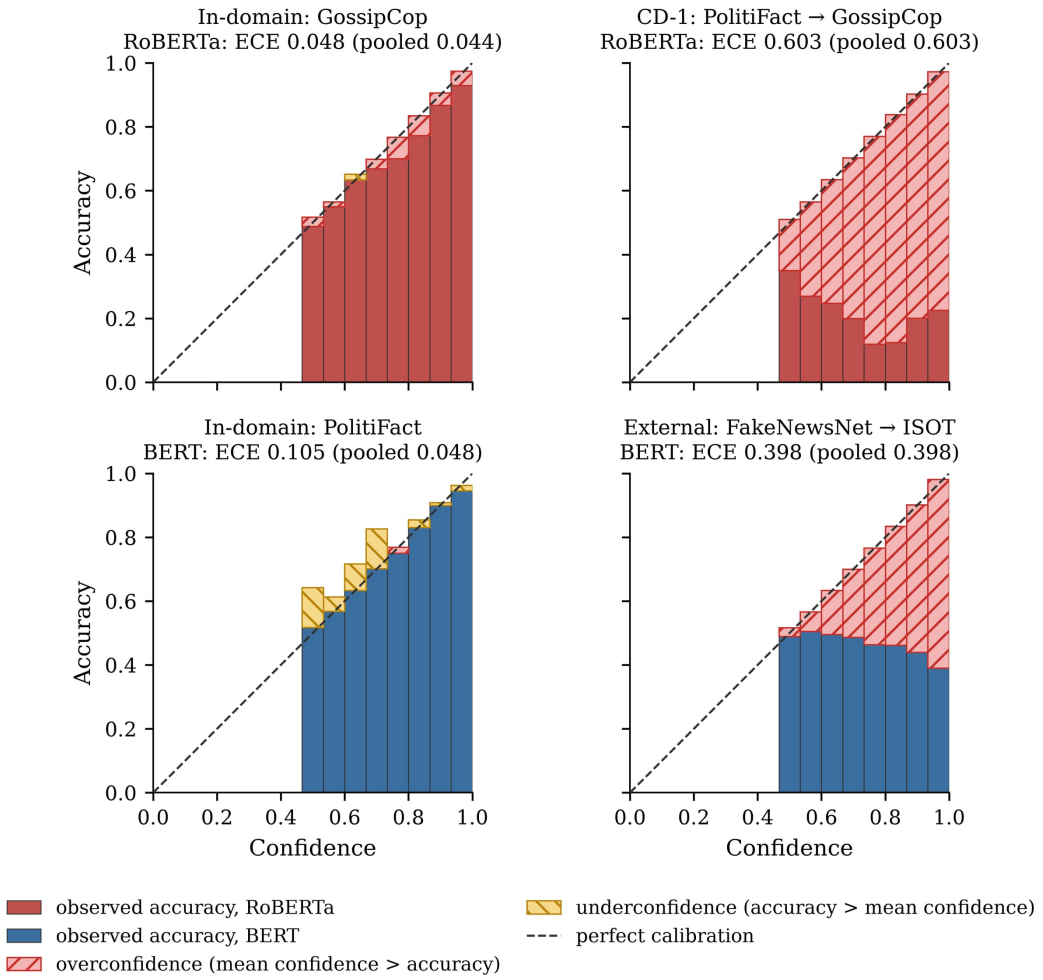


Figure 6. Reliability diagrams under in-domain and domain-shift conditions, illustrating calibration degradation under domain shift.

Table 9. Calibration under domain shift. Expected calibration error (15 equal-width bins) and Brier score; mean ± standard deviation over five seeds.

Setting	Metric	BERT	RoBERTa	ELECTRA	DistilRoBERTa
PolitiFact (in-domain)	ECE	0.105 ± 0.032	0.137 ± 0.058	0.115 ± 0.037	0.215 ± 0.090
	Brier	0.151 ± 0.025	0.156 ± 0.047	0.143 ± 0.010	0.193 ± 0.049
GossipCop (in-domain)	ECE	0.077 ± 0.043	0.048 ± 0.031	0.044 ± 0.033	0.057 ± 0.041
	Brier	0.123 ± 0.011	0.108 ± 0.013	0.110 ± 0.008	0.113 ± 0.010
PolitiFact → GossipCop	ECE	0.419 ± 0.119	0.603 ± 0.249	0.427 ± 0.058	0.390 ± 0.166
	Brier	0.425 ± 0.078	0.595 ± 0.199	0.419 ± 0.038	0.364 ± 0.137
GossipCop → PolitiFact	ECE	0.379 ± 0.104	0.224 ± 0.115	0.240 ± 0.107	0.247 ± 0.076
	Brier	0.393 ± 0.080	0.272 ± 0.079	0.307 ± 0.069	0.298 ± 0.041
FakeNewsNet → ISOT	ECE	0.398 ± 0.162	0.227 ± 0.121	0.296 ± 0.069	0.244 ± 0.043
	Brier	0.443 ± 0.114	0.306 ± 0.063	0.346 ± 0.041	0.303 ± 0.026
Mean ECE, in-domain		0.091	0.092	0.079	0.136
Mean ECE, under shift		0.399	0.352	0.321	0.294
ECE increase factor		4.4×	3.8×	4.0×	2.2×

The correction improved performance in all four models for the transfer from PolitiFact to GossipCop, where substantial Fake-class overprediction was observed, with Macro-F1 improvements ranging from 0.038 to 0.159. The effect is smaller in external evaluation but remains positive. Under in-domain conditions and in the GossipCop → PolitiFact direction, the correction generally reduces performance. The average correction effect across the

evaluated conditions was approximately zero at +0.002, with improvements concentrated in settings characterized by substantial Fake-class overprediction rather than class-prior mismatch alone.

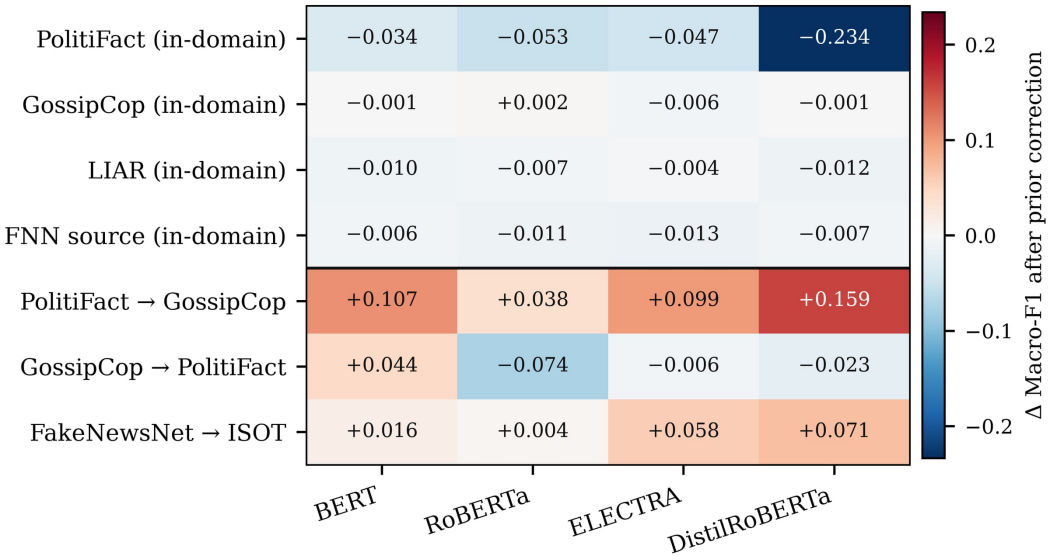


Figure 7. Effect of post-hoc class-weight compensation at inference across evaluation conditions, illustrating the direction-dependent variation in Macro-F1.

Table 10. Effect of post-hoc class-weight compensation at inference. Positive values indicate improvement in Macro-F1 relative to the argmax decision rule.

Setting	BERT	RoBERTa	ELECTRA	DistilRoBERTa
PolitiFact (in-domain)	−0.034	−0.053	−0.047	−0.234
GossipCop (in-domain)	−0.001	+0.002	−0.006	−0.001
LIAR (in-domain)	−0.010	−0.007	−0.004	−0.012
FakeNewsNet source (in-domain)	−0.006	−0.011	−0.013	−0.007
PolitiFact → GossipCop	+0.107	+0.038	+0.099	+0.159
GossipCop → PolitiFact	+0.044	−0.074	−0.006	−0.023
FakeNewsNet → ISOT	+0.016	+0.004	+0.058	+0.071

The results of the diagnostic analysis can be seen in Table 11. The loss decomposition, threshold-sensitivity analyses, uncertainty estimates, and exploratory statistical tests associated with Table 11 are detailed in Supplementary Tables S1, S2, S9 and S10. Additional class-conditional distribution diagnostics are provided in Supplementary Tables S4–S8. In the transfer from PolitiFact to GossipCop, cross-fitted threshold adjustment yields an improvement of 0.315 Macro-F1 points, corresponding to approximately 63% of the observed total Macro-F1 loss. The recoverable difference is much smaller in the reverse transfer from GossipCop to PolitiFact (0.003) and in the transfer from FakeNewsNet to ISOT (0.060). The directional pattern is also found under balanced accuracy and after removing the training-time class weighting. The post-hoc correction effect is related to Fake-class overprediction, as can be seen from a Spearman correlation coefficient of 0.71. This explains why the average effect of the post-hoc correction is near-zero across heterogeneous evaluation conditions. The hierarchical bootstrap analysis also supports this directional difference. In the transfer from PolitiFact to GossipCop, the cross-fitted threshold-sensitive share is 0.63 with a 95% confidence interval of 0.54 to 0.77. The corresponding cross-fitted performance difference is 0.315 Macro-F1 points with a 95% confidence interval of 0.285 to 0.342. The difference is positive and also statistically significant after Holm correction. In the reverse transfer from

GossipCop to PolitiFact, no statistically significant recoverable difference is found. These results characterize threshold sensitivity and do not uniquely identify its causal contribution to performance loss.

Table 11. Quantification of threshold-sensitive performance loss and class-conditional distribution shift.

Diagnostic	PF → GC	GC → PF	FNN → ISOT
Total Macro-F1 loss	0.500	0.234	0.241
Cross-fitted threshold difference	0.315	0.003	0.060
Relative threshold-sensitive share	63%	≈1%	25%
Real-class score distribution difference (KS statistic)	0.787	0.617	0.686
Real-class domain AUROC (TF-IDF)	0.995	0.995	0.998

Class-conditional diagnostics showed that the transfer asymmetry could not be adequately explained by class-prior mismatch alone. In the transfer from PolitiFact to GossipCop, the mean Fake-class logit for true Real instances shifted from −1.20 in the source domain to +1.19 in the target domain, with a Kolmogorov–Smirnov statistic of 0.787. Within-class domain classifiers also strongly distinguished PolitiFact from GossipCop, achieving AUROC values of 0.976 for Fake instances and 0.995 for Real instances. Representation-level maximum mean discrepancy analyses provided further evidence of substantial domain differences. Moreover, synthetic label-shift controls produced substantially smaller class-prior estimation errors than actual cross-domain transfer. These findings are inconsistent with a pure label-shift explanation and indicate substantial class-conditional distribution changes, with possible changes in feature–label relationships. The corresponding class-conditional score-distribution, lexical domain-discrimination, encoder-representation, synthetic prior-shift, and exploratory feature–label diagnostics are presented in Supplementary Tables S4, S5, S6, S7 and S8, respectively.

4.3. Statistical Comparison Analysis

Performance differences among the four models were evaluated to determine whether they remained statistically significant beyond variability introduced by random seeds.

The Friedman test was conducted across 35 blocks, corresponding to 7 evaluation scenarios and 5 random seeds, with the four models treated as the comparison groups. The test indicated a significant overall difference among the models, $\chi^2(3) = 14.79$, $p = 0.0020$. Secondary full-target evaluation sets were excluded from the block structure. The mean ranks of the models and the results of the subsequent pairwise comparisons are reported in Table 12 and Figure 8.

Table 12. Statistical comparison of the four models using the Friedman test and Holm-corrected pairwise Wilcoxon signed-rank tests. Lower mean rank indicates better performance, and negative effect sizes favor the second model in each pair.The downward arrow (↓) indicates that lower mean rank values are preferable.

Model	Mean Rank ↓	Pairwise Comparison	p (Raw)	p (Holm)	Decision	Effect Size r	Magnitude
DistilRoBERTa	2.13	BERT vs. DistilRoBERTa	0.0096	0.0576	Fail to reject	−0.509	Large
RoBERTa	2.14	BERT vs. RoBERTa	0.0118	0.0590	Fail to reject	−0.483	Medium
ELECTRA	2.57	BERT vs. ELECTRA	0.0182	0.0730	Fail to reject	−0.454	Medium
		ELECTRA vs. RoBERTa	0.1037	0.3111	Fail to reject	−0.318	Medium
BERT	3.16	DistilRoBERTa vs. ELECTRA	0.1898	0.3796	Fail to reject	+0.257	Small
		DistilRoBERTa vs. RoBERTa	0.8020	0.8020	Fail to reject	−0.051	Negligible

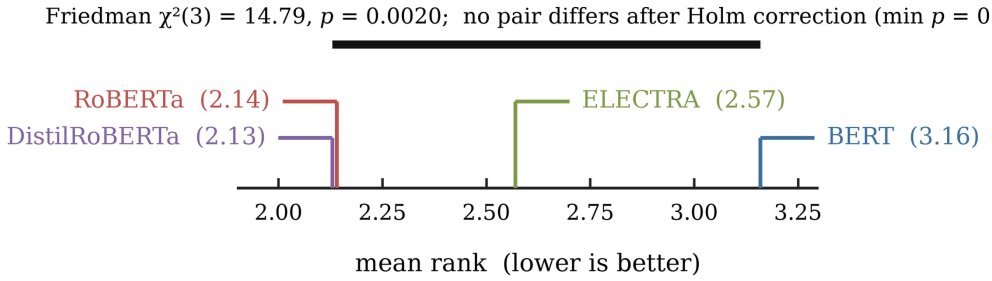


Figure 8. Mean ranks of the four encoders across the 35 Friedman blocks. Although the omnibus test is significant, no pairwise comparison remains significant after Holm correction.

Following the significant Friedman test, pairwise Wilcoxon signed-rank tests were performed with Holm correction for multiple comparisons. After correction, none of the model pairs differed significantly at the $\alpha = 0.05$ level; the lowest adjusted p-value was 0.058. Although rank-biserial effect sizes were moderate to large in some comparisons, these results do not support a general superiority claim for any individual model. Negative effect sizes indicate an advantage for the second model in the corresponding pairwise comparison.

4.4. Ablation Analysis

The proposed interpretation of cross-domain degradation was further examined through two controlled ablation experiments targeting plausible alternative explanations: training-time class weighting and differences in source-domain training size. The results are summarized in Table 13. Additional threshold-sensitivity analyses for the ablation conditions are reported in Supplementary Tables S1, S2 and S9.

Table 13. Ablation results for class weighting and matched training size. Macro-F1 values are reported as mean ± standard deviation over five seeds.

Ablation	Evaluation Set	Model	Main Matrix	Ablated	Δ Macro-F1
A1: unweighted loss	PolitiFact (in-domain)	BERT	0.770 ± 0.038	0.774 ± 0.032	+0.004
		RoBERTa	0.782 ± 0.041	0.820 ± 0.027	+0.038
		ELECTRA	0.756 ± 0.041	0.747 ± 0.050	−0.010
		DistilRoBERTa	0.778 ± 0.030	0.787 ± 0.040	+0.009
	CD-1 PolitiFact → GossipCop	BERT	0.324 ± 0.076	0.351 ± 0.034	+0.027
		RoBERTa	0.217 ± 0.073	0.193 ± 0.008	−0.024
		ELECTRA	0.311 ± 0.035	0.395 ± 0.076	+0.084
		DistilRoBERTa	0.235 ± 0.030	0.236 ± 0.031	+0.001
A2: matched training size	GossipCop (in-domain)	BERT	0.772 ± 0.005	0.652 ± 0.022	−0.120
		RoBERTa	0.784 ± 0.018	0.695 ± 0.012	−0.089
		ELECTRA	0.780 ± 0.007	0.645 ± 0.017	−0.135
		DistilRoBERTa	0.780 ± 0.011	0.688 ± 0.008	−0.091
	CD-2 GossipCop → PolitiFact	BERT	0.465 ± 0.075	0.462 ± 0.058	−0.003
		RoBERTa	0.609 ± 0.059	0.590 ± 0.043	−0.019
		ELECTRA	0.533 ± 0.041	0.498 ± 0.030	−0.035
		DistilRoBERTa	0.571 ± 0.035	0.604 ± 0.022	+0.033

In the first ablation (A1), class weighting was removed from the training objective while all other settings were kept unchanged. This modification increased the mean in-domain Macro-F1 on PolitiFact by 0.010 and the mean cross-domain Macro-F1 in the transfer from PolitiFact to GossipCop by 0.022. Neither difference was statistically significant, with $p = 0.295$ and $p = 0.154$, respectively. Removing class weighting nevertheless altered the decision boundary in the expected direction: Fake-class precision increased by 0.028–0.099 across the four models, whereas recall decreased by 0.016–0.080. However, this change did not eliminate the cross-domain collapse. Even without class weighting, the models still

classified 73.2–97.2% of GossipCop test instances as Fake, compared with 81.8–94.7% under the weighted setting.

The second ablation (A2) examined whether the directional asymmetry could be attributed to the much larger GossipCop training set. The GossipCop training split was therefore reduced from 8823 to 335 instances, matching the size of the PolitiFact training split while preserving the original class ratio. As expected, this substantially reduced in-domain GossipCop performance, with a mean Macro-F1 change of -0.109 ($p < 0.001$). In contrast, CD-2 performance remained essentially unchanged, with a mean change of -0.006 ($p = 0.869$). Under the matched-size condition, CD-2 Macro-F1 values remained between 0.462 and 0.604 and therefore did not approach the substantially lower CD-1 range of 0.217–0.324. Because in-domain performance decreased while transfer performance remained comparatively stable, the corresponding retention rates increased to 70.8–87.7%.

Taken together, these ablations indicate that the observed transfer asymmetry cannot be explained solely by training-time class weighting or by differences in source-domain training-set size. However, these ablations do not exclude other domain-specific factors, including differences in text length, truncation severity, lexical distribution, and topic composition.

Additional input-length analyses were conducted to assess whether truncation could explain the observed directional asymmetry. For DistilRoBERTa, increasing the maximum sequence length from 128 to 512 tokens improved in-domain PolitiFact Macro-F1 from 0.779 to 0.841, whereas PolitiFact-to-GossipCop transfer Macro-F1 remained low (0.198–0.239). Across the three input lengths, approximately 93–97% of GossipCop examples were predicted as Fake, despite an actual Fake prevalence of 20%. The cross-fitted share of the observed Macro-F1 loss remained between 0.63 and 0.69 (Supplementary Table S13). In the reverse direction, cross-fitted differences were comparatively small, with confidence intervals including zero. A complementary single-seed chunked-inference analysis across four encoders showed a similar directional pattern, including in a subset of documents identified as not requiring truncation under the BERT tokenizer. These findings suggest that truncation alone is insufficient to explain the observed directional asymmetry, although input length affects absolute predictive performance. Detailed results are provided in Supplementary Tables S13 and S14.

5. Discussion

Three key findings can be derived from the present study: First, source-based shortcuts exist in measurable magnitude. Second, cross-domain transfer is highly asymmetric; in other words, transfer from source to target is much more effective than transfer in the reverse direction. Third, calibration drift, defined here as a deterioration in calibration accompanying performance degradation, has been quantitatively measured. These findings can be used to qualify the traditional interpretation of failure on out-of-domain instances: such failure may involve decision-rule and/or calibration effects in addition to representational degradation.

Shortcut validation shows that a classifier based on only 53 terms in the ISOT dataset achieved a Macro-F1 score of 0.994. This value suggests that a substantial portion of the high performance reported on this dataset may rely on source-signature detection rather than content understanding. This finding is consistent with previous warnings regarding artificialities arising from dataset construction and aligns with the observation that even relatively small distribution shifts can lead to measurable bias. It is also noteworthy that the density of shortcuts varies greatly across datasets: while 77.8% of ISOT records contain at least one shortcut term, this proportion is 0.4% in LIAR. The comparatively lower in-domain scores reported here may partly reflect the stricter duplicate and shortcut-cleaning protocol; therefore, direct numerical comparisons with studies using different cleaning procedures should be interpreted cautiously.

A more detailed analysis of the cross-domain degradation beyond the Macro-F1–AUROC divergence is presented. For the transfer from PolitiFact to GossipCop, cross-fitted threshold adjustment recovers approximately 63% of the observed Macro-F1 loss. This is in contrast to the reverse transfer from GossipCop to PolitiFact where only approximately 1% of the observed Macro-F1 loss can be recovered. For the external transfer, approximately 25% of the observed Macro-F1 loss can be recovered by cross-fitted threshold adjustment. These results show that while discrimination also deteriorates in the transfer from PolitiFact to GossipCop, substantial performance can still be recovered by threshold adjustment. Note that these are threshold-sensitive performance differences under a specified diagnostic procedure and do not uniquely identify causal effects. Bootstrap confidence intervals as well as adjusted statistical tests support direction-dependent threshold sensitivity. However, the small PolitiFact test set as well as the dependence among model–seed observations severely limit the uncertainty estimates. In summary, the additional diagnostic analysis supports the central hypothesis that threshold-sensitive degradation contributes to cross-domain performance loss in a direction-dependent manner. However, the negligible difference in the reverse transfer from GossipCop to PolitiFact indicates that this mechanism does not explain performance degradation equally across all transfer conditions.

Although PolitiFact and GossipCop have different class-prior probabilities (34% vs. 20%), the presence of label shift is not established solely based on class-prior mismatch. To establish pure label shift, one must assume that the class-conditional input distribution remains constant across domains. This assumption was tested using score-level, text-based, and representation-level diagnostic tools. The most direct explanation for the observed prediction imbalance when transferring from PolitiFact to GossipCop is the marked displacement of Real-class scores toward Fake predictions. Therefore, the evidence supports both class-prior differences and significant class-conditional distribution changes, which may include changes in feature–label relationships. As such, post-hoc class weight compensation can serve as a diagnostic tool for evaluating the sensitivity of a decision rule. However, its directional effects and near-zero average benefit do not suggest that it should be used as a general-purpose performance improvement method.

Calibration deterioration was observed across alternative measures. However, ECE estimates varied across methods, particularly for the 73-instance PolitiFact test set. The reported 2.2- to 4.4-fold increases therefore apply only to the original 15-bin ECE formulation.

The fact that the transfer values reported in Table 7 fall below those of many published cross-domain results is not a methodological weakness but a direct consequence of protocol differences. As detailed in Section 2.3, studies reporting high cross-domain performance typically access unlabeled data from the target domain or include the target domain in a multi-domain training mix; in this study, however, no data from the target domain—whether labeled, unlabeled, or for model selection—was used. To ensure comparability across evaluation protocols, the F1 score for the positive class is also reported. For the "Fake" class, the F1 score derived from the precision and recall values in Table 7 ranges from 0.337 to 0.345 for the PolitiFact → GossipCop direction, from 0.369 to 0.469 for the GossipCop → PolitiFact direction, and from 0.459 to 0.606 for external evaluation. These values are consistent with the range reported in the literature for protocols without access to the target domain. Additionally, the in-domain ceiling measured for retention is lower than the ceiling in studies using less stringent duplicate and shortcut-cleaning procedures; therefore, the rates reported here are calculated based on a lower but more accurate reference.

Three practical implications emerge from these findings. First, decision-rule calibration should be examined before attributing transfer failure solely to representation degradation. Second, in-domain performance alone is insufficient for model selection. Differences among

models are small in Table 6 but reach 0.144 Macro-F1 points in Table 7. Third, source-based shortcut auditing should be included in fake news detection evaluations.

Several limitations should be considered when interpreting these findings. The maximum sequence length was limited to 128 tokens due to computational constraints, resulting in the truncation of 88.4% of the documents in the GossipCop training set and 77.0% of those in the PolitiFact set; the reported absolute performance values may underestimate the performance attainable with longer input contexts. Although the same maximum sequence length was applied to all models, architecture- and tokenizer-specific sensitivity to truncation cannot be excluded. Moreover, PolitiFact documents are approximately twice as long as GossipCop documents on average (6710 vs. 3200 characters), while substantial truncation occurs in both domains. Consequently, the observed transfer asymmetry should not be attributed exclusively to class-prior or decision-rule mismatch. Differences in document length, the amount and location of information removed by truncation, and other domain-specific characteristics may also contribute to the directional effect. While the additional input-length and chunking analyses further confirmed the directional pattern, they did not completely disentangle the effects of truncation, domain-specific content and other distributional differences. Long-context retraining was restricted to DistilRoBERTa with three seeds, while chunking was analyzed at inference time with one seed. The FakeNewsNet subset used in this study is smaller than the canonical dataset. The PolitiFact training set contains only 335 examples, which may cause the observed Macro-F1 standard deviations of 0.030–0.041. Five random seeds were therefore used to assess training variability, and the data partitions remained identical across runs in order to attribute the observed variation to training stochasticity rather than differences in the dataset splits. DeBERTa-v3 could not be included in the evaluation because it did not converge under the common fixed learning rate specified in the original design; this observation is consistent with prior evidence that fine-tuning hyperparameters can affect architectures differently and limits the generalizability of the comparison to the four models evaluated here. Although the Friedman test was significant, no pairwise comparison remained significant after Holm correction, indicating that the present experimental design does not support a definitive claim of superiority for any individual architecture.

6. Conclusions

This study investigated performance degradation in cross-domain fake news detection. A controlled experimental design was used to examine data leakage and source-based shortcut signals. Four Transformer encoders were evaluated under in-domain, cross-domain, and external evaluation conditions with five random seeds, and a total of 80 training runs were conducted.

The findings support the central hypothesis that cross-domain Macro-F1 degradation reflects both reduced discriminative performance and a threshold-sensitive component whose magnitude varies with transfer direction. In the PolitiFact-to-GossipCop transfer, cross-fitted threshold adjustment yielded a Macro-F1 improvement of 0.315, with a 95% hierarchical bootstrap confidence interval of 0.285 to 0.342. This corresponds to approximately 63% of the observed performance loss, with a 95% confidence interval of 54% to 77%. The improvement was supported by an exploratory Holm-adjusted test, with an adjusted p-value of 1.9×10^{-5}. In contrast, the reverse transfer showed an improvement of only 0.003, with a 95% confidence interval of -0.053 to 0.016 and no statistically significant improvement. The external FakeNewsNet-to-ISOT evaluation showed an intermediate threshold-sensitive share of 25%, with an improvement of 0.060 and a 95% confidence interval of 0.047 to 0.076. Class-conditional analyses further indicated that differences in class prevalence alone could not explain the observed degradation. Together, these results identify threshold sensitivity

as an important, but not universal, contributor to cross-domain performance loss. The estimated improvements reflect diagnostic threshold sensitivity rather than causal effects or gains achievable without target-domain labels.

In terms of evaluation validity, the fact that a classifier based on only 53 source-related terms achieved a Macro-F1 of 0.994 on ISOT revealed that superficial source signals can significantly affect reported performance. Therefore, it is important to systematically report duplicate and shortcut checks in fake news detection studies.

Furthermore, it was observed that model rankings varied depending on the direction of transfer. Models with similar in-domain performance differed substantially under domain shift. However, no pairwise model comparison remained statistically significant after multiple-comparison correction. This result does not support the general superiority of a single Transformer architecture.

Future work will need to extend the input-length robustness analysis to more encoder architectures, more training seeds and more long-context modeling strategies. Also, post-hoc class-weight compensation should be compared to other methods of label-shift adaptation and calibration. Shortcut removal should be studied in more detail using controlled ablations. Token-level signals should be studied using SHAP-based explanations. Finally, the evaluation framework should be applied to more languages and multilingual datasets to see if the observed calibration drift and decision-rule misalignment generalize to other domains.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app162111026/s1. Table S1: Macro-F1 accounting and cross-fitted threshold sensitivity; Table S2: Uncertainty and class-prevalence-independent sensitivity check; Table S3: Compensation effect stratified by predicted Fake-class imbalance; Table S4: Class-conditional prediction-score distribution diagnostics; Table S5: Domain discrimination within each true class using lexical features; Table S6: Representation-level diagnostics for PF versus GC; Table S7: Synthetic prior-shift controls compared with actual transfer; Table S8: Exploratory evidence of feature–label association differences; Table S9: Cross-fitted mechanism estimates with 95% hierarchical bootstrap confidence intervals; Table S10: Exploratory Holm-adjusted one-sided tests for mechanism-related quantities; Table S11: Shifted-to-in-domain calibration deterioration ratios with 95% confidence intervals; Table S12: Absolute ranges across seven ECE-type estimators, by evaluation condition; Table S13: Input-length sensitivity of DistilRoBERTa in bidirectional cross-domain evaluation; Table S14: Chunked-inference sensitivity across Transformer encoders.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available within the Supplementary Materials.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (GPT-5.6 Sol) and Claude (Opus 5 and Sonnet 5) for purposes such as text refinement, language editing, and structural improvement. The author reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 109. [CrossRef]
2. Silva, A.; Luo, L.; Karunasekera, S.; Leckie, C. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 557–565. [CrossRef]

3. Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; Li, J. MDFEND: Multi-domain Fake News Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event, 1–5 November 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 3343–3347. [CrossRef]
4. Bozarth, L.; Budak, C. Toward a Better Performance Evaluation Framework for Fake News Classification. *Proc. Int. AAAI Conf. Web Soc. Media* **2020**, *14*, 60–71. [CrossRef]
5. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. *arXiv* **2017**, arXiv:1706.04599. [CrossRef]
6. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv* **2019**, arXiv:1906.02530. [CrossRef]
7. Blanco-Fernández, Y.; Otero-Vizoso, J.; Gil-Solla, A.; García-Duque, J. Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models. *Appl. Sci.* **2024**, *14*, 9729. [CrossRef]
8. Moisi, E.; Mihalca, B.; Coman, S.; Pater, A.; Popescu, D. Romanian Fake News Detection Using Machine Learning and Transformer-Based Approaches. *Appl. Sci.* **2024**, *14*, 11825. [CrossRef]
9. Bucos, M.; Drăgulescu, B. Enhancing Fake News Detection in Romanian Using Transformer-Based Back Translation Augmentation. *Appl. Sci.* **2023**, *13*, 13207. [CrossRef]
10. Alyoubi, S.; Kalkatawi, M.; Abukhodair, F. The Detection of Fake News in Arabic Tweets Using Deep Learning. *Appl. Sci.* **2023**, *13*, 8209. [CrossRef]
11. Cavus, N.; Oktekin, B.; Goksu, M. A Cloud-Based Sentiment Analysis System with a BERT Algorithm for Fake News on Twitter. *Appl. Sci.* **2025**, *15*, 11046. [CrossRef]
12. Alarfaj, F.; Khan, H.; Naz, A.; Almusallam, N. A Real-Time Large Language Model Framework with Attention and Embedding Representations for Misinformation Detection. *Eng. Appl. Artif. Intell.* **2026**, *164*, 113304. [CrossRef]
13. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
14. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
15. Clark, K.; Luong, M.-T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv* **2020**, arXiv:2003.10555. [CrossRef]
16. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108. [CrossRef]
17. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv* **2020**, arXiv:2006.03654. [CrossRef]
18. Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; et al. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 2526–2547. [CrossRef]
19. Yelmen, I.; Gunes, A.; Zontul, M. Multi-Class Document Classification Using Lexical Ontology-Based Deep Learning. *Appl. Sci.* **2023**, *13*, 6139. [CrossRef]
20. Rout, J.; Mishra, M.; Saikia, M. Towards Reliable Fake News Detection: Enhanced Attention-Based Transformer Model. *J. Cybersecur. Priv.* **2025**, *5*, 43. [CrossRef]
21. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? In *Chinese Computational Linguistics*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; pp. 194–206. [CrossRef]
22. Kocun, G.; Yelmen, I.; Zontul, M.; Yildirim, F. Early and Accurate Diagnosis of Ectopic Pregnancy: A SHAP-Based Explainable Machine Learning Approach. *IEEE Access* **2026**, *14*, 80511–80530. [CrossRef]
23. Hong, S.; Wang, X.; Zhang, C.; Wang, J.; Duan, P.; Wang, Y. AIGC Video Detection Based on the Fusion of Spatial-Frequency-Optical Flow Multimodal Features. *J. Syst. Eng. Electron.* **2026**, 1–15. [CrossRef]
24. Jiang, K.; Wang, Q.; An, Z.; Wang, Z.; Zhang, C.; Lin, C.-W. Mutual Retinex: Combining Transformer and CNN for Image Enhancement. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 2240–2252. [CrossRef]
25. Blackledge, C.; Atapour-Abarghouei, A. Transforming Fake News: Robust Generalisable News Classification Using Transformers. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021*; IEEE: New York, NY, USA, 2021; pp. 3960–3968. [CrossRef]
26. Elangovan, A.; He, J.; Verspoor, K. Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1325–1335. [CrossRef]

27. Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
28. Bissoto, A.; Barata, C.; Valle, E.; Avila, S. Even Small Correlation and Diversity Shifts Pose Dataset-Bias Issues. *Pattern Recognit. Lett.* **2024**, *179*, 87–93. [CrossRef]
29. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [CrossRef] [PubMed]
30. 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, 30 July–4 August 2017*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 422–426. [CrossRef]
31. Ahmed, H.; Traore, I.; Saad, S. Detecting Opinion Spams and Fake News Using Text Classification. *Secur. Priv.* **2018**, *1*, e9. [CrossRef]
32. Zou, K.; Xu, Y.; Yu, J.; Pan, Y.; Yin, J.; Lai, H. GAFD-CC: Global-Aware Feature Decoupling with Confidence Calibration for Out-of-Distribution Detection. *Neural Netw.* **2026**, *199*, 108728. [CrossRef] [PubMed]
33. Raghavan, D.; Bhandari, S.; Banerjee, J.; Bandyopadhyay, S. Out-of-Distribution Detection in Text Using Statistical Techniques. *Discov. Data* **2025**, *3*, 46. [CrossRef]
34. He, Y.; Peng, L.; Zhang, Y.; Weng, J.; Li, S.; Luo, Z. Long-Tailed Out-of-Distribution Detection: Prioritizing Attention to Tail. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 3446–3454. [CrossRef]
35. Saerens, M.; Latinne, P.; Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Comput.* **2002**, *14*, 21–41. [CrossRef] [PubMed]
36. Nguyen, A.; Tran, L.; Tong, A.; Nguyen, T.; Tran, T. CASUAL: Conditional Support Alignment for Domain Adaptation with Label Shift. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 19668–19676. [CrossRef]
37. Naeini, M.P.; Cooper, G.; Hauskrecht, M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proc. AAAI Conf. Artif. Intell.* **2015**, *29*, 2901–2907. [CrossRef]
38. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
39. Jin, Y.; Choi, M.; Verma, G.; Wang, J.; Kumar, S. MM-SOC: Benchmarking Multimodal Large Language Models in Social Media Platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 6192–6210. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.