

Enhancing Trust in Online Reviews: A Cross-Domain Fake Review Detection Framework Using Generative AI

Richa Gupta¹, Indu Kashyap², Vinita Jindal^{3*}

^{1,3}Keshav Mahavidyalaya, University of Delhi, Delhi, India

²Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India

Abstract – Online reviews influence consumer decisions in e-commerce but are threatened by deceptive content. Detecting these fake reviews across domains remains challenging due to limited labelled data, class imbalance, and domain-specific bias. This work proposes CORAL, a unified retrieval-augmented framework for cross-domain fake review detection that mitigates source-domain bias through unified masking and generates target-aware counterfactuals via retrieval grounding with orthogonal alignment. Unlike prior domain-adaptation approaches, CORAL enables zero-shot classification without manual annotation. Evaluated across twelve cross-domain tasks on benchmark e-commerce datasets, CORAL achieved average gains of 21.74 % in accuracy and 24.25 % in F1-score over strong baselines while improving trustworthiness and robustness. It demonstrated stability across random seeds, resilience to moderate noise, reduced hallucination, and competitive runtime efficiency. Additionally, CORAL enabled the annotation of a novel healthcare review dataset, addressing scarcity of labelled data.

Keywords – Consumer behaviour, cross-domain, e-commerce, fake review detection, generative AI, RAG, transfer learning.

I. INTRODUCTION

Online reviews strongly influence consumer purchasing decisions and business reputation in digital marketplaces. As reliance on reviews has increased, deceptive content designed to manipulate consumer perception and competitive standing has prevailed. Fake reviews mislead users, undermine fair competition, and erode trust in online platforms. Detecting such reviews has therefore become a critical research challenge, particularly with the emergence of generative models capable of producing human-like text. Recent surveys indicate a sharp decline in trust in online reviews, with consumer confidence dropping from 79 % in 2020 to 42 % in 2025 [1], while only 19 % of users report confidence in identifying fake reviews. These trends highlight both the increasing sophistication of fake reviews and the limitations of manual detection.

Fake Review Detection (FRD) has been widely studied using machine learning techniques [2] that analyse linguistic [3], stylistic [4], visual [5], and behavioural cues [6]. While these approaches perform well within a single domain, their effectiveness declines substantially when applied across

domains due to domain-specific vocabulary and contextual differences. This limitation is further compounded by the scarcity of fully annotated datasets in socially significant domains such as healthcare and education [7], where manual annotation is costly, unreliable, and increasingly impractical.

Cross-domain Fake Review Detection (CDFRD) addresses these challenges by transferring knowledge from labelled source domains to unlabelled target domains in a zero-shot setting. However, existing cross-domain approaches often rely solely on source-domain information or ignore target-domain context, limiting generalization. Moreover, as regulatory bodies and e-commerce platforms increasingly emphasise transparency, accountability, and trustworthy AI, scalable automated review moderation has become a strategic necessity rather than merely a technical concern.

To overcome these limitations, this paper proposes CORAL, a unified framework for Cross-domain fake review detection using Orthogonal, Retrieval-Augmented counterfactual Learning. CORAL suppresses source-domain bias through unified masking and generates target-domain-aware counterfactual reviews using Gemini Pro 2.5 combined with retrieval-augmented generation (RAG). These counterfactuals preserve authenticity-related cues while enabling automated zero-shot labelling without human intervention.

The main contributions of this study are listed below.

1. A novel, retrieval-augmented framework CORAL for addressing CDFRD that integrates masking, generative AI and target-aware retrieval.
2. A fully automated zero-shot labelling approach that eliminates manual annotation in low-resource domains.
3. Annotation of a novel healthcare review dataset, demonstrating CORAL's scalability and applicability in real-world e-commerce and service ecosystems.

II. RELATED WORK

Most FRD research focuses on single-domain settings, limiting generalization in cross-domain scenarios. Li et al. [8] pioneered cross-domain FRD using SVM and SAGE but evaluated only two domain pairs. Ren and Ji [9] proposed a

*Corresponding author's e-mail: vjindal@keshav.du.ac.in
Article received 2026-02-17; accepted 2026-05-15

bidirectional GRNN with attention, trained solely on hotel reviews. Cagnina et al. [10] used character n-grams and LIWC features, achieving modest cross-domain performance. Sánchez-Junquera et al. [11] introduced domain adaptation via co-occurring entropy and feature mismatch yet remained dataset-specific. Ren et al. [12] combined GloVe, TF-IDF, CNN, and DistilBERT but evaluated only limited pairs. Gupta et al. [13] introduced COAT with vocabulary masking and counterfactuals but did not address class imbalance. These limitations restrict the real-world applicability of existing CDFRD methods, where reviews span multiple product and service domains and labelled data remain scarce.

Counterfactuals model how reviews would appear if written in a different domain while preserving their authenticity labels. Previously explored in sentiment analysis [14] and fake news detection [15], they have been shown to reduce domain bias. However, their application in FRD remains limited, often relying on heuristic or static transformations that lack retrieval-based incorporation of target-domain semantics. This gap motivates retrieval-augmented counterfactual learning for CDFRD.

Zero-shot labelling enables classification without annotated target data using clustering [16], embedding similarity [17], and semantic distance optimisation [18]. Although LLMs enable zero-shot learning, retrieval-augmented counterfactual generation remains largely unexplored, especially for large-scale e-commerce platforms, where rapid onboarding of new domains and services make manual annotation infeasible.

III. METHODOLOGY FOR THE PROPOSED FRAMEWORK CORAL

This section presents CORAL's four-phase architecture with auxiliary module. It is illustrated in Fig. 1, with the pseudocode given in Appendix A for reproducibility.

A. Problem Definition

Let the source domain be $D_S = \{(r_i^S, l_i^S)\}_{i=1}^{N_S}$, with labelled reviews r_i^S and corresponding labels $l_i^S \in \{\text{fake}, \text{real}\}$. Let the target domain be $D_T = \{(r_j^T)\}_{j=1}^{N_T}$ containing unlabelled reviews. The objective of CDFRD is to learn a decision function $F: D_S \rightarrow D_T$ that predicts labels $\hat{l}_j^T$ to target-domain reviews r_j^T in a zero-shot setting.

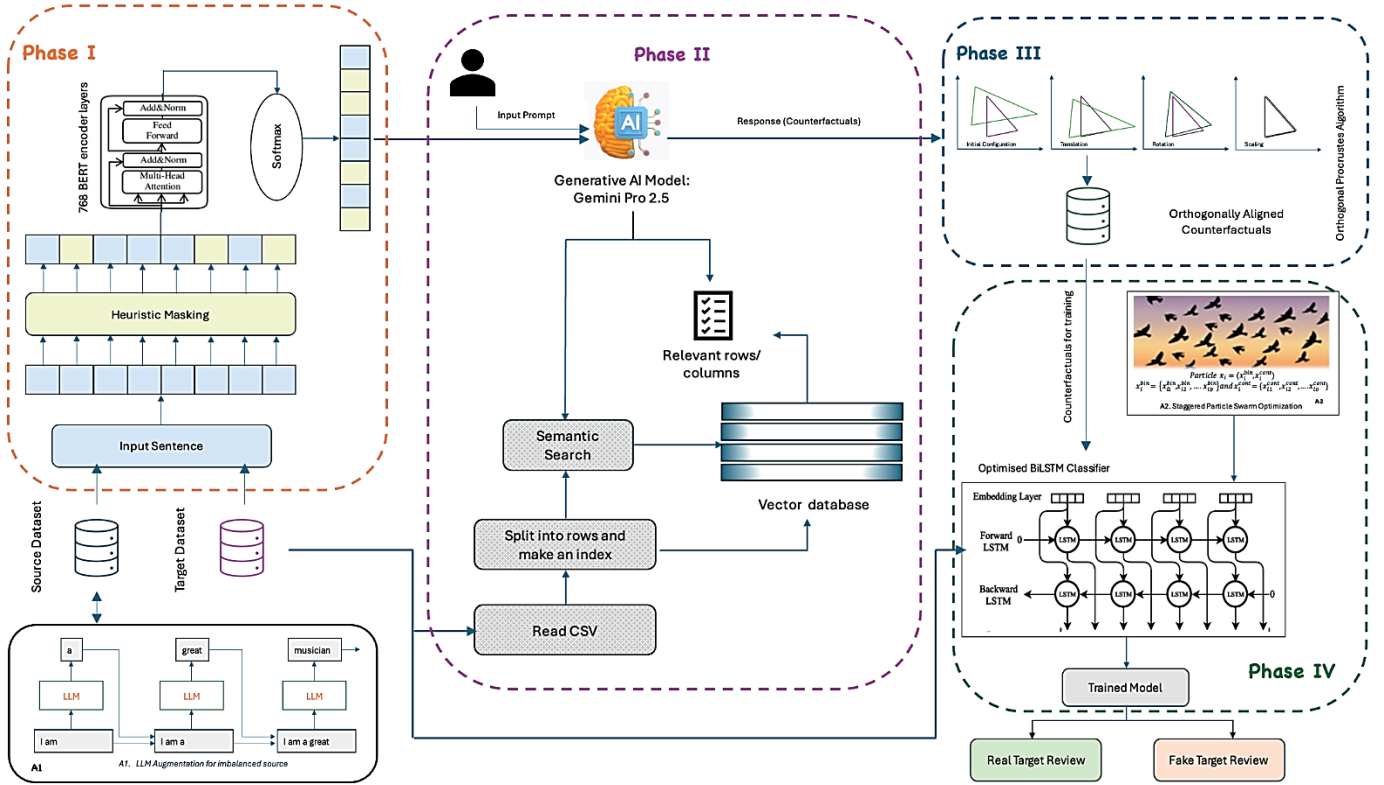


Fig. 1. Architecture of the proposed framework CORAL.

B. Phase I: Unified Masking (UM)

The first phase of CORAL employs Unified Masking (UM), integrating heuristics and context, to suppress source-domain-specific tokens. Originally introduced in COAT [13], UM is refined in CORAL to enhance domain generalization by reducing reliance on domain-biased cues. The unified masking score $M(w, D_S, D_T)$ (1), integrates the heuristic domain-

affinity score $M_1(w, D_S, D_T)$ (2) and the contextual attention-based significance $M_2(w, D_S)$ (3), balanced by a weighting factor $\delta \in [0, 1]$.

$$M(w, D_S, D_T) = \delta \cdot M_1(w, D_S, D_T) + (1 - \delta) \cdot M_2(w, D_S) \quad (1)$$

$$M_1(w, D_S, D_T) = \rho(w, D_S) - \rho(w, D_T) \quad (2)$$

where

$$\rho(w, d) = P(d|w) \cdot \left(1 - \frac{H(d|w)}{\log N}\right) \quad (2a)$$

$$H(w|d) = -\sum_{i=1}^n P(d_i|w) \cdot \log_2 P(d_i|w) \quad (2b)$$

$$M_2(w, D_S) = \sum_{i=1}^N \left(\frac{\exp(\alpha_i)}{\sum_{j=1}^N \exp(\alpha_j)} \right) \quad (3)$$

where

$$\alpha_i = v_i^T \cdot c \quad (3a)$$

Tokens whose scores exceed a threshold τ are replaced with a [MASK] token (4).

$$\text{Mask}(word) = \begin{cases} 1, & \text{if } M(word, D_S, D_T) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

This process removes domain-biased vocabulary while preserving contextual semantics. Empirically, $\delta = 0.5$ and $\tau = 0.8$ produced stable performance across datasets. These masked reviews are then forwarded to Phase II.

C. Phase II: Counterfactual Generation

Phase II automates counterfactual generation by integrating Gemini Pro 2.5 with Retrieval-Augmented Generation (RAG). Gemini is selected for its strong semantic generalization, cost-latency efficiency, and prompt controllability, while RAG ensures grounding in target-domain context and reduces hallucination. Masked, domain-neutral reviews from Phase I are provided as input to Gemini, which generates counterfactuals using retrieved target-domain context and a user-defined prompt. RAG embeds and indexes the target data, retrieving the most relevant chunks to guide contextual infilling of masked tokens. This process aligns vocabulary, tone, and semantics with the target domain while preserving original labels. The resulting counterfactuals retain source intent and structure but reflect target-domain linguistic features, thereby bridging cross-domain gaps without manual intervention. These counterfactuals are then passed onto Phase III.

D. Phase III: Orthogonal Alignment of Counterfactuals

Phase III applies Orthogonal Procrustes Alignment (OPA) to align generated counterfactuals with target-domain embeddings while preserving their geometric and semantic structure. Adopted from COAT [13], OPA learns an orthogonal transformation matrix $\mathbb{Q}^*$ by minimizing the Frobenius norm between $\mathbb{Q}X$ and Y , i.e., $\mathbb{Q}^* = \arg \min_{\mathbb{Q}} \|\mathbb{Q}X - Y\|_F$ for counterfactual embeddings X and target-domain embeddings Y . The transformed embeddings $X' = \mathbb{Q}^* X$ reduce residual source bias. The combined process of Phases II and III is summarised in (5).

$$\mathbb{C} := \text{OPA}\{\text{GeminiPro}[\mathcal{R}'_i, \text{prompt}, \text{RAG}(\mathcal{D}_T)]\} \quad (5)$$

where $\mathbb{C}$ denotes the aligned counterfactuals and $\mathcal{R}'_i$ are the masked source reviews. This alignment is crucial in e-

commerce, where vocabulary varies significantly across product and service domains.

E. Phase IV: Classification and Prediction

In the final phase, a computationally efficient Bi-LSTM is trained on Gemini-generated counterfactuals, each inheriting the corresponding source label. Since authenticity-related cues like linguistic complexity and stylistic patterns, remain largely preserved during counterfactual generation, they provide reliable labelled data for zero-shot classification of target reviews as fake or real without manual intervention.

F. Auxiliary Module: Handling Imbalance in Source Dataset

This auxiliary module is invoked only when the source dataset is imbalanced, i.e., when the real reviews significantly outnumber the fake ones. To ensure fair learning, the minority class is augmented using GPT-3.5. Domain-relevant reviews are generated via contextual augmentation, where selected tokens are replaced or expanded while preserving semantic integrity. To reduce hallucination, generated outputs are filtered using trust scores. Selected samples form a balanced source D'_S . This lightweight augmentation avoids costly fine-tuning, maintains domain consistency, and reduces majority-class bias.

Together, these phases and auxiliary module constitute the unified transfer learning framework CORAL, which effectively bridges domain gaps, enabling robust CDFRD.

IV. EXPERIMENTAL SETUP

CORAL was evaluated on two fully labelled benchmark groups, each spanning three domains (Table I), widely used in prior studies [19]. Group I reflects realistic cross-domain e-commerce: Amazon Online Products dataset [20], Deceptive Opinion Spam Corpus [8], and Yelp Labelled dataset [21], drawn from independent platforms with high linguistic diversity. Group II comprises hotel, restaurant, and doctor reviews compiled by Li et al. [8]. Although multi-domain Group II datasets share a unified collection methodology and exhibit possible stylistic overlap, they remain a standard benchmark for CDFRD.

TABLE I
BENCHMARK DATASETS USED FOR EXPERIMENTATION

GROUP	DOMAIN	DATASET NAME	CLASS DISTRIBUTION
I	Online Products	Amazon Fake Review Dataset [20]	20K fake + 20K real
	Hotel	Deceptive Opinion Spam Corpus [8]	800 fake + 800 real
	Restaurant	Yelp Labelled Dataset [21]	36 885 fake + 321 271 real
II	Restaurant	Deceptive opinion spam corpus [8]	357 fake + 200 real
	Doctor		200 fake + 200 real
	Hotel		800 fake + 800 real

Each group yields six source-target tasks, totalling twelve experiments. CORAL was evaluated on all tasks (Table II).

TABLE II
TWELVE TASKS FOR CORAL EVALUATION

GROUP I (SOURCE → TARGET) TASKS	
Online products → Rest. (P→R)	Hotels → Restaurant (H→R)
Online products → Hotels (P→H)	Hotels → Online Products (H→P)
Rest. → Online Products (R→P)	Restaurant → Hotels (R→H)
GROUP II (SOURCE → TARGET) TASKS	
Hotels → Restaurant (H→R)	Doctor → Restaurant (D→R)
Restaurant → Hotels (R→H)	Doctor → Hotels (D→H)
Restaurant → Doctor (R→D)	Hotels → Doctor (H→D)

A. Experiments in Phase I: Unified Masking

Unified masking was applied across all domain pairs, suppressing source-specific tokens to produce domain-neutral reviews while retaining original labels. Sensitivity analysis on the weighting factor δ and masking threshold τ assessed robustness (Section V.E(d)). Table III presents Hotel→Online Products example depicting this phase.

TABLE III
REPRESENTATIVE EXAMPLE

Original Hotel review	The Talbot hotel is an elegant place to take you wife or husband for a weekend getaway. This personally yours romantic package was fabulous.
After Unified masking	The [MASK][MASK] is an [MASK][MASK] to take your wife or husband for a [MASK][MASK]. The personally yours [MASK] package was fabulous.

B. Experiments in Phase II: Counterfactual Generation

This phase used Gemini Pro 2.5 with RAG to generate high-quality counterfactuals from masked reviews. Target datasets were indexed using sentence embeddings in ChromaDB. Given a masked review and a structured user prompt, Gemini retrieved the top-k chunks via cosine similarity to provide contextual guidance during mask filling. For instance, in P→R, masked Product reviews were used as input while Restaurant reviews provided target context. Prompts followed a fixed structure with domain-specific examples. Gemini then generated counterfactuals that preserved original intent while adopting target's vocabulary and style. For comparison, GPT-3.5 was also evaluated but produced less coherent outputs. All experiments were implemented in Python on Colab Pro (T4 GPU). Table IV showcases the directional prompts used for counterfactual generation.

TABLE IV
REPRESENTATIVE PROMPT USED FOR P→R TASK

ORIGINAL MASKED REVIEW	PROMPT FOR CORAL	GENERATED COUNTERFACTUAL
This [MASK] works great, and the [MASK] is very easy to use. I bought it last week	"Assume the role of a professional restaurant-review writer. Your task is to transform the masked passage into a coherent restaurant review while preserving the original sentence structure and sentiment polarity.	★★★★★ This place works great and the service is very easy to appreciate. I visited last week and it has been extremely satisfying. Definitely

and it has been extremely [MASK]. Definitely a good [MASK] for everyday use.	Transformation requirements: 1. Replace masked tokens with vocabulary typical of restaurant reviews such as food, ambience, service etc. 2. Maintain the intent and tone of the original text. 3. Ensure the output matches the style of reviews found on platforms such as Yelp or TripAdvisor. 4. Use natural restaurant-domain phrasing, avoiding exaggerated embellishments. Guiding Example: If the input is "This device works great and I use it everyday," your output should resemble "This dish tastes great and I would order it again." Now transform the passage accordingly."	a good spot for everyday dining. OR This restaurant works wonderfully and the ambience is very easy to settle into. I came here last week and had an extremely pleasant experience. A great choice for a casual daily meal.
--	--	---

C. Experiments in Phase III: Alignment of Counterfactuals

The impact of orthogonal alignment was evaluated via an ablation study in which OPA was removed from CORAL. Figure 2 shows consistent declines in accuracy, confirming its key role in reducing residual domain shift and improving cross-domain generalization.

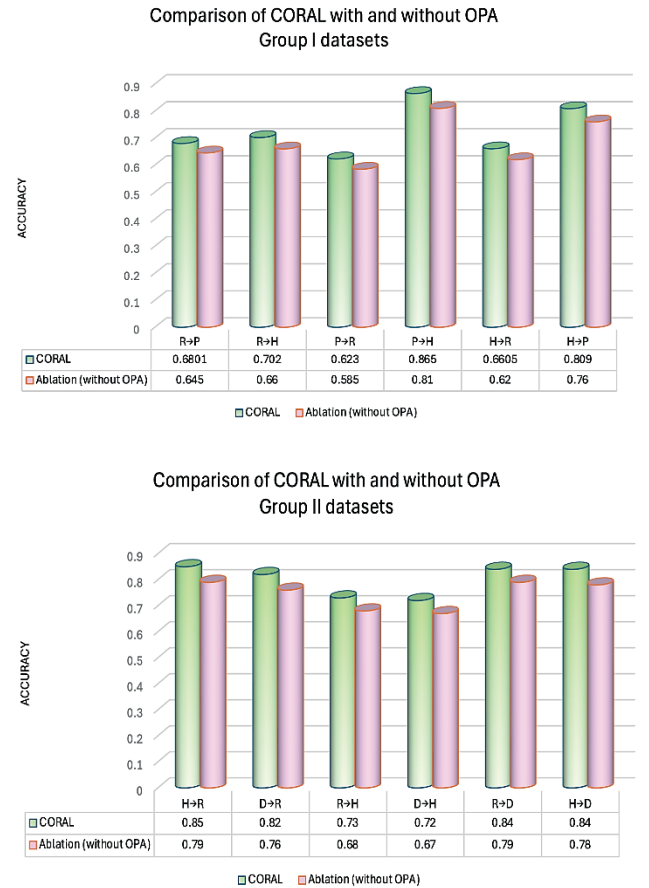


Fig. 2. Comparison of CORAL with and without OPA in terms of accuracy.

D. Experiments in Phase IV: Classification and Prediction

In this final phase, a Bi-LSTM (128 units, 20 % dropout) was trained on aligned counterfactuals from Phase III using inherited source labels. It predicted labels for unseen target

reviews, whose true labels remained hidden during training. Predictions were then compared with original target labels to compute per-task accuracy, subsequently averaged across tasks.

V. RESULTS

A. Comparison of CORAL with Baselines

CORAL's performance was evaluated by comparing predicted and true target labels and benchmarking against six baselines: NoDoM (No Domain Modification for direct transfer), RMTF (Random Masking Target Fill which randomly masks 10 % source tokens), GPT-4 (prompt-based counterfactuals), RoBERTa and DeBERTa (MLM infilling without retrieval), and Gemini Pro 2.5 without RAG (generation without target-domain grounding). Table V reports accuracy, F1-score, and trustworthiness. Accuracy reflected

overall performance; NoDoM and RMTF performed poorly due to domain mismatch and random substitutions, while other baselines improved performance through pretrained models but lacked target grounding. CORAL achieved higher F1-scores, indicating more balanced precision and recall. Additionally, trustworthiness [22] was computed as:

$$\text{TRUST} - \text{SCORE} = \frac{P+R+F1+Acc}{4} \quad (6)$$

with all metrics normalized to [0,1]. Higher scores denoted more reliable cross-domain detection, whereas lower scores indicated retrieval or classification weaknesses. Trust levels were categorised as high (≥ 0.65), medium (0.50–0.65) and low (< 0.50). CORAL attained medium-to-high trust levels, whereas most baselines remained in the low-to-medium range.

TABLE V
COMPARISON OF BASELINES WITH CORAL IN TERMS OF ACCURACY, F1 SCORE, AND TRUST-SCORE

Src → Tgt PAIR		P → R	H → R	R → P	H → P	P → H	R → H	AVERAGE
Framework	Metrics (%)							
NoDoM (No Domain Modification)	Accuracy	44.00	52.50	50.00	52.00	50.00	55.00	50.59
	F1 Score	40.11	45.47	50.21	44.84	55.03	50.71	47.73
	Trust-Score	41.39	46.24	50.86	45.94	55.34	51.7	48.58
RMTF (Random Masking Target Fill)	Accuracy	48.00	50.50	54.00	50.00	48.50	52.00	50.50 (−0.09)
	F1 Score	40.75	48.79	53.86	47.75	58.85	53.8	50.64 (+2.91)
	Trust-Score	42.33	49.97	55.21	48.26	59.70	57.00	52.08 (+3.5)
GPT4	Accuracy	56.00	53.00	59.50	51.00	53.00	50.00	53.75 (+3.16)
	F1 Score	42.19	50.67	55.86	49.21	59.89	55.27	52.19 (+4.76)
	Trust-Score	42.84	52.03	56.59	49.40	61.89	57.93	53.45 (+4.87)
RoBERTa	Accuracy	47.12	53.50	51.00	52.00	54.00	52.50	51.60 (+1.01)
	F1 Score	41.70	50.80	56.00	48.95	60.20	55.80	52.24 (+4.51)
	Trust-Score	43.10	52.40	56.80	50.10	62.00	58.50	53.81 (+5.23)
DeBERTa	Accuracy	48.50	54.50	53.00	53.00	55.00	56.00	53.34 (+2.75)
	F1 Score	42.40	51.30	56.90	49.70	61.00	56.20	52.91 (+5.18)
	Trust-Score	44.00	53.00	57.50	50.90	62.80	59.70	54.65 (+6.07)
Gemini without RAG	Accuracy	51.00	53.00	57.00	54.00	56.00	60.00	55.17 (+4.58)
	F1 Score	42.35	51.50	56.11	49.46	61.08	56.75	52.88 (+5.45)
	Trust-Score	44.54	53.56	57.69	50.85	62.92	60.96	55.09 (+6.51)
CORAL (Proposed Framework)	Accuracy	62.30	66.05	68.01	80.90	86.50	70.20	72.33 (+21.74)
	F1 Score	66.10	70.00	60.10	80.09	85.50	70.04	71.98 (+24.25)
	Trust-Score	57.67	66.08	60.62	59.12	66.03	66.35	62.65 (+14.07)

To assess robustness and reproducibility, CORAL was evaluated across six tasks using five random seeds. Performance showed low variance (Fig. 3). Results were reported as mean $\pm$ standard deviation, with 95 % confidence intervals (7) computed using Student's t-distribution to account for small sample sizes with unknown population variance.

$$CI = \bar{x} \pm t_{n-1, 0.975} \frac{s}{\sqrt{n}} \quad (7)$$

where $\bar{x}$ denotes the mean performance across seeds, s is the standard deviation, and $n = 5$ is the number of seeds.

To further understand CORAL's behaviour under domain shift, representative false positives and false negatives are shown in Appendix B, see Table XI.

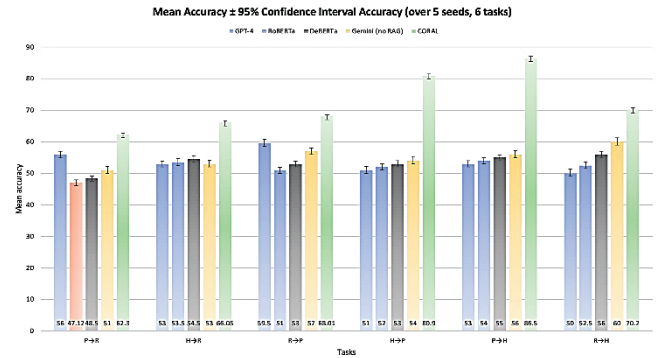


Fig. 3. Mean accuracy over 5 seeds and 6 tasks.

B. Comparison of CORAL with Prior Research Works

Prior cross-domain FRD studies mainly reported results on Group II datasets, with no comparable results available for Group I. Therefore, CORAL was compared only on Group II tasks. The results are presented in Table VI. The compared, prior works included SAGE [8], GRNN [9], n-gram + SVM [10], SOR-XGB [23], MASK-BERT [24], DV-MA [11], MFE [12], and COAT [13], spanning Bayesian, neural, feature-based, masking, and counterfactual alignment approaches.

TABLE VI
CORAL WITH PRIOR WORKS ON GROUP II DATASETS

SRC→TGT	H→R	R→H	H→D	D→H	R→D	D→R
METHOD	ACCURACY					
SAGE [8]	0.7700	-	0.6470	-	-	-
GRNN [9]	0.8170	-	0.5570	-	-	-
n-gram+SVM [10]	0.6525	0.6585	0.6397	0.5744	0.6397	0.5746
SOR-XGB [23]	-	-	0.6906	-	0.6906	-
MASK-BERT [24]	0.8158	0.6595	0.7250	0.6891	0.7300	0.8000
MFE [12]	0.8300	-	0.7800	-	-	-
COAT [13]	0.8351	0.6620	0.8102	0.7018	0.8225	0.8094
CORAL	0.8500	0.8200	0.8400	0.7200	0.8400	0.8000
Method	F1-SCORE					
SAGE [8]	0.7840	-	0.6280	-	-	-
GRNN [9]	0.8060	-	0.6590	-	-	-
n-gram+SVM [10]	0.6400	0.6600	0.5000	0.4200	0.5000	0.5700
DV-MA [11]	0.7900	0.7100	0.6600	0.6700	0.6400	0.6200
MASK-BERT [24]	0.8728	0.5882	0.7650	0.6449	0.6142	0.8849
MFE [12]	0.8330	-	0.7540	-	-	-
COAT [13]	0.8740	0.6625	0.8034	0.6443	0.8213	0.8851
CORAL	0.8900	0.9000	0.8300	0.6600	0.8400	0.9000

Performance gains observed in Table VI stemmed from:

- context-aware masking with dynamic retrieval: Unlike DV-MA's static masking, CORAL retrieved and integrated target context to generate more realistic counterfactuals;
- stronger domain adaptation: By embedding target knowledge into counterfactuals, CORAL adapted more effectively than GRNN and MFE, which lacked contextual augmentation;
- richer semantic representations: Unlike feature-based methods (SAGE, n-gram+SVM), CORAL used semantically aligned counterfactuals to improve cross-domain generalization.

C. Runtime Efficiency and Scalability Analysis

Table VII compares CORAL's runtime efficiency with baselines on Group I tasks, covering counterfactual generation (per 100 reviews) and classification time. CORAL was faster than GPT-4 and COAT and was competitive with Gemini (No RAG) and DeBERTa. For example, in $P \rightarrow R$ task, CORAL took 30.53 min versus 40.05 (GPT-4) and 221.5 (COAT), demonstrating clear efficiency gains from retrieval-guided generation and orthogonal alignment. Additionally, CORAL completed classification within 49–62 min, comparable to or faster than GPT-4, DeBERTa, Gemini (No RAG), and COAT. This improvement was attributed to domain-aligned counterfactuals that reduced ambiguity and reprocessing. For e-commerce applications, CORAL balanced computational cost and detection quality, with a one-time

generation overhead. Although the starred cases in Table VII increased generation time due to imbalance handling, they improved classification accuracy and reliability justifying, the additional overhead.

TABLE VII
RUNTIME EFFICIENCY COMPARISON WITH BASELINES

SRC → TGT	GPT4	DeBERTa	GEMINI (No RAG)	COAT [13]	CORAL
COUNTERFACTUAL GENERATION (PER 100 COUNTERFACTUALS)					
$P \rightarrow R$	40.05	42.15	40.05	221.5	30.53
$H \rightarrow R$	46.58	46.90	46.58	229.2	38.44
$R \rightarrow P$	43.12	41.80	43.12	234.7	56.93*
$H \rightarrow P$	48.03	48.60	48.03	245.1	39.11
$P \rightarrow H$	52.20	52.95	52.20	251.3	44.21
$R \rightarrow H$	47.22	49.10	47.22	259.1	58.24*
CLASSIFICATION TIME					
$P \rightarrow R$	57	72	75	54	55
$H \rightarrow R$	60	65	55	51	51
$R \rightarrow P$	59	70	70	75	77
$H \rightarrow P$	55	62	56	60	61
$P \rightarrow H$	58.30	65	65	60	62
$R \rightarrow H$	54.50	60	52.30	49	51

Note: *includes LLM augmentation using auxiliary module to address imbalance in Restaurant dataset.

D. Hallucination in Generative AI

Hallucination refers to the generation of fluent yet factually incorrect content, which threatens e-commerce review analysis. CORAL mitigated this issue through RAG and structured prompting and was quantitatively assessed using trust scores (Table V) and perplexity (Fig. 4). Perplexity measures a model's confidence in token prediction [25], [26] (8).

$$\text{Perplexity} = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i) \right) \quad (8)$$

where N is the number of tokens, and $P(w_i)$ is the predicted probability of the i^{th} token. Lower perplexity indicates more coherent, less hallucinated outputs. Figure 4 shows that CORAL achieved the lowest perplexity across tasks, indicating minimum hallucination and more trustworthy generation. This reliability is critical for consumer decisions and business reputation.

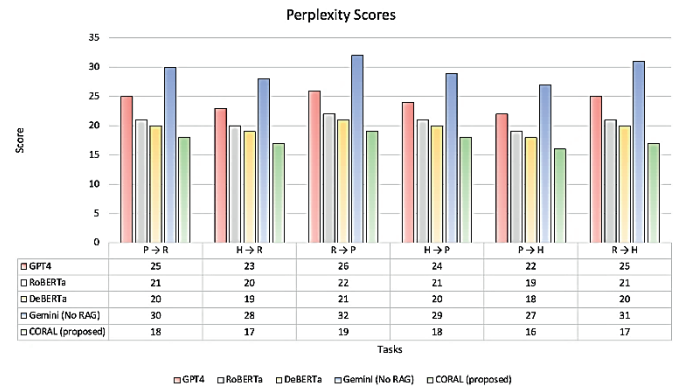


Fig. 4. Perplexity scores of compared frameworks across tasks.

E. Sensitivity Analysis

This section presents a sensitivity analysis to assess CORAL's robustness.

Impact of the Size of Unlabelled Target Corpus

To assess sensitivity to target data size, the unlabelled target corpus was subsampled at {1 %, 5 %, 10 %, 25 %, 50 %, 100 %}, with accuracy and F1-scores reported in Fig. 5.

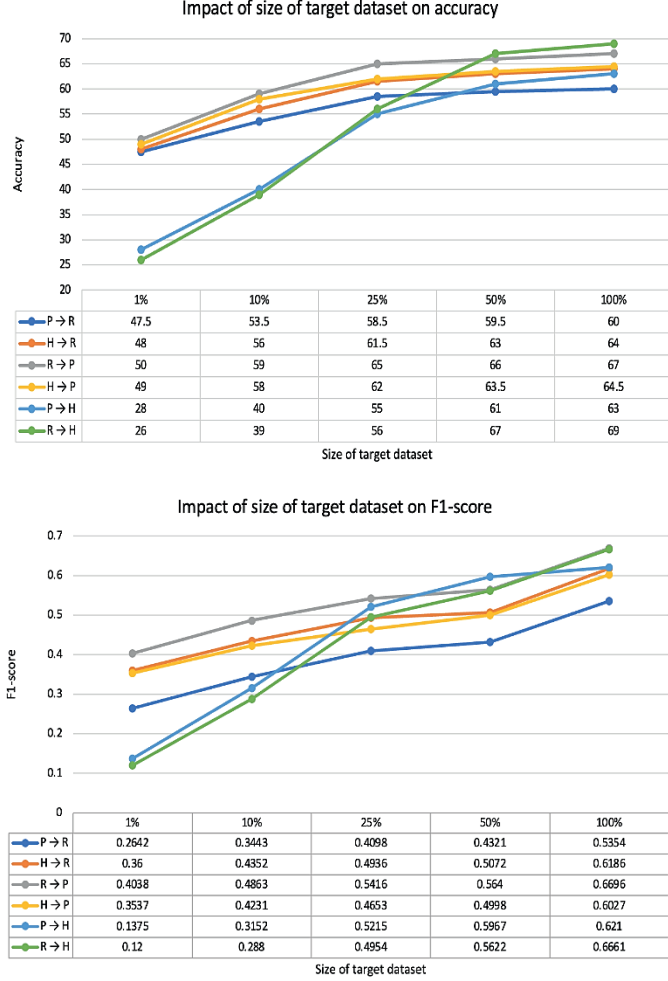


Fig. 5. Impact of size of target dataset on Accuracy and F1-score.

Figure 5 shows that for larger target datasets, accuracy and F1-score improved gradually, as even small subsets sufficed for RAG. In contrast, for smaller H dataset, it dropped sharply at very low sampling (e.g., only 16 samples at 1 %) but recovered as more data became available. Although CORAL consistently benefited from increased target data, its effectiveness declined when the target was extremely small. Empirically, stable results emerged with ~200–400 target samples (≈ 15 – 25 % of smaller and ≥ 1 % of larger datasets). Below this threshold, retrieval indexing was insufficient to provide reliable grounding.

Impact of the Noise in Unlabelled Target Corpus

Controlled noise was introduced in the unlabelled target corpus at levels of {0 %, 10 %, 30 %, 50 %}. Figure 6 reports the resulting accuracy and F1-scores. At 10 % noise, CORAL

exhibited only minor performance degradation, indicating robustness to mild contamination. However, as target noise increased to 30 % and 50 %, accuracy declined noticeably, particularly for smaller Hotel dataset where noisy samples constituted a larger proportion of the data. F1-scores degraded more rapidly for smaller targets, decreasing steadily with increasing noise. Overall, CORAL remained stable up to 30 % noise, beyond which performance dropped sharply.

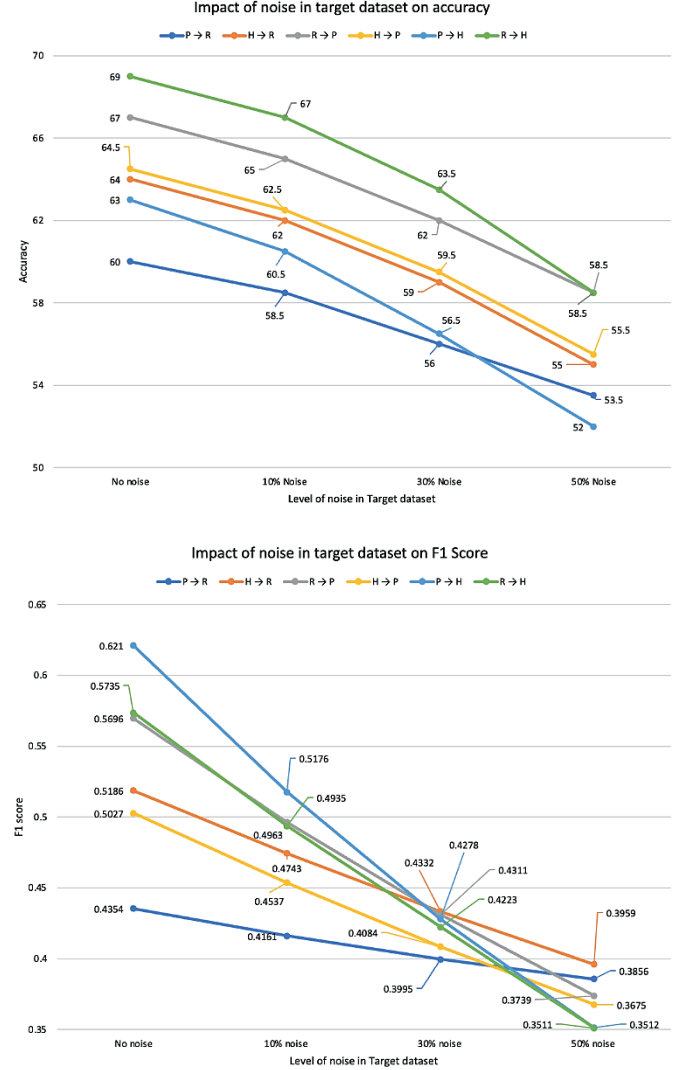


Fig. 6. Impact of noise in target dataset on Accuracy and F1-score.

Sensitivity to Retriever Hyperparameters

To assess sensitivity to retrieval settings, three retriever models (BM25, MiniLM, mpnet) and the number of retrieved passages (top- $k = 1, 3, 5, 10, 15$), were varied. Heatmaps in Fig. 7 illustrate that accuracy improved with stronger retrievers and increasing k , saturating around $k \approx 5$ – 10 . In comparison, perplexity decreased as k increased, indicating reduced hallucinations due to richer target-domain grounding. Smaller targets (e.g., Hotel) benefited less from larger k . Although accuracy continued to rise at $k = 15$, the gains over $k = 5$ – 10 were marginal, suggesting an optimal trade-off between retrieval depth and efficiency. While Mpnet achieved the

highest accuracy, it was computational costly; whereas the lighter MiniLM offered comparable efficiency. Beyond the optimal k , the retriever started pulling in less relevant or redundant chunks, increasing computational cost and

reintroducing noise despite minor accuracy gains. Overall, a moderate retrieval depth ($k \approx 5-10$) balanced accuracy, trustworthiness and runtime.

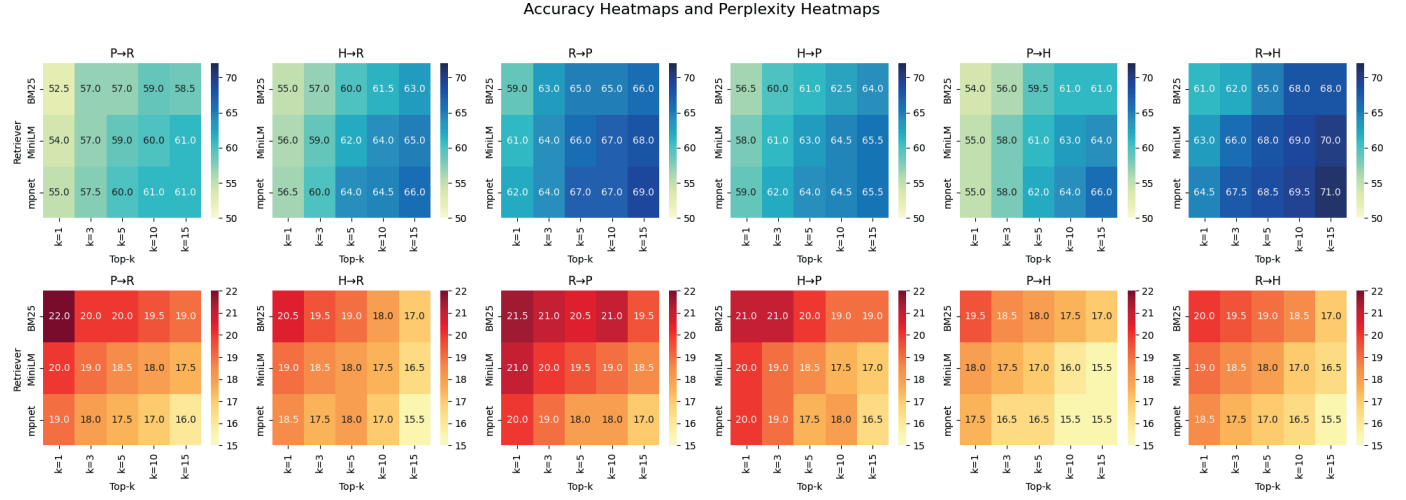


Fig. 7. Heatmaps showing sensitivity to retriever hyperparameters.

Sensitivity to Weighing Factor δ and the Masking Threshold τ

Robustness of unified masking was evaluated via a sensitivity analysis of the weighing factor δ and the masking threshold τ . The parameter $\delta \in [0,1]$ controlled the contribution of heuristic and contextual components, while τ controlled masking intensity. Several representative values of δ were examined with a fixed τ , and vice-versa. Figures 8(a) and 8(b) exhibit only modest performance variations, peaking at $\delta = 0.5$ and $\tau = 0.8$. This configuration achieved an effective balance between suppressing source bias and preserving semantic context for counterfactual generation. Accordingly, these values were adopted in CORAL for all experiments.

F. Ethical Considerations

CORAL enhances CDFRD via automated counterfactual generation, but also introduces potential misuse risks, as it can produce fluent, label-consistent reviews, that could be exploited to generate deceptive content. To prevent this, safeguards such as (i) controlled access to generator, (ii) secure authentication, (iii) digital watermarking for forensic traceability, and (iv) adherence to ethical AI standards are recommended. Transparency and accountability are essential when deploying counterfactual generation in operational settings. Additional caution is needed in sensitive domains such as healthcare and education. Recommendations include: (i) using public datasets only, (ii) enforcing data de-identification, (iii) validating outputs with domain experts, and (iv) conducting periodic bias and safety audits to ensure CORAL's trustworthy, compliant, and socially responsible deployment in real-world e-commerce and service platforms.

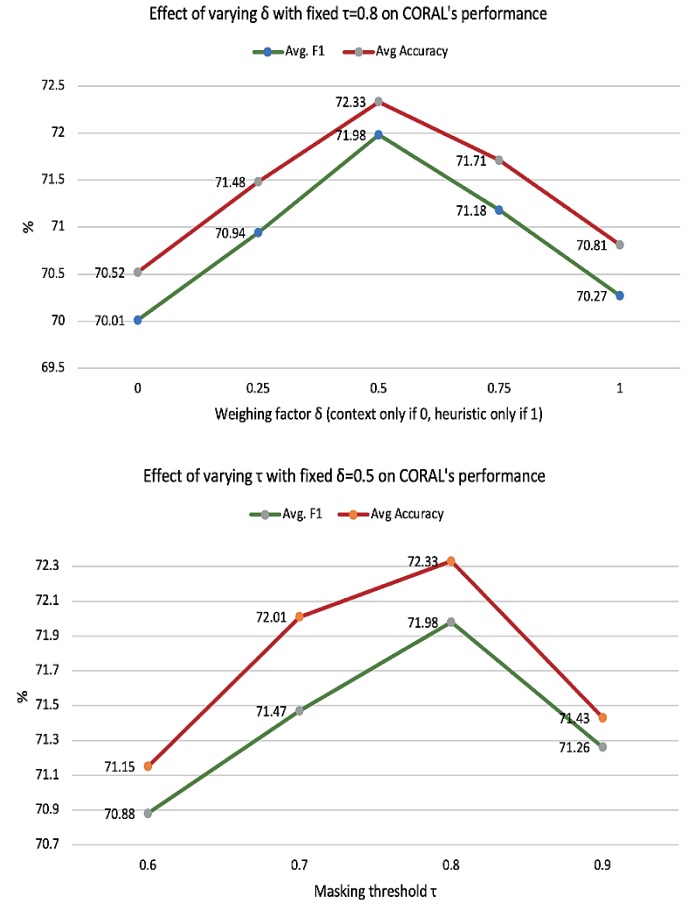


Fig. 8. Sensitivity to weighing factor and Sensitivity to masking threshold.

VI. ANNOTATION OF A NOVEL HEALTHCARE DATASET

A key motivation of this research is scarcity of fully annotated review datasets in domains such as healthcare. To address this gap, a novel healthcare review dataset was created and zero-shot annotated using CORAL. Reviews scraped from prominent Indian hospital websites formed the unlabelled target dataset. The annotation procedure consists of three steps:

- Data collection and preprocessing;
- Source dataset selection, and
- Annotation using CORAL.

A. Methodology for Annotating the Scraped Dataset

Data Collection and Preprocessing

Healthcare reviews were scraped using instant Data Scraper tool (v1.2.0) [27]. After removing non-verified reviews manually, 2011 verified reviews formed target dataset. Proper nouns, including hospital and doctor names, were anonymised using generic tokens as part of pre-processing. A RAG database was built from the cleaned corpus. Table VIII presents samples of anonymised healthcare reviews.

TABLE VIII

SAMPLE OF HEALTHCARE REVIEWS SCRAPED FROM THE WEB

S. No.	REVIEW
1	I had a wonderful experience at X Hospital. The doctors were very attentive, explained everything in detail, and the nursing staff was extremely caring throughout my stay. The facilities were clean and well-maintained. I felt safe and well looked after.
2	The emergency department team acted quickly when I was admitted for chest pain. Within minutes I was diagnosed and treated. Special thanks to Dr. Y and the ICU nurses for their professionalism and compassion.

Source Dataset Selection

Healthcare reviews exhibit service-oriented characteristics resembling hospitality domain. Following Panda & Levitan [28], cross-dataset cosine similarity was computed where Healthcare-Hotel showed the highest similarity (0.0743), versus Healthcare-Restaurant (0.0543) and Healthcare-Product (0.0326). Accordingly, the Hotel dataset was selected as source.

Annotation Using CORAL

CORAL was applied twice on the Hotel→Healthcare pair for robustness. Source reviews were masked, healthcare-guided counterfactuals were generated, and target reviews were automatically labelled as fake or real.

B. Results and Discussion

Label Consistency

Table IX shows highly consistent predictions across runs, with approximately 46 % fake and 54 % real reviews. The negligible variation between iterations confirms the stability and absence of systematic bias in CORAL's annotations.

TABLE IX

ANNOTATION OF THE UNLABELLED DATASET

ITERATION	SOURCE→TARGET	NO. OF PREDICTED FAKE	NO. OF PREDICTED REAL	TOTAL REVIEWS
1	Hotel →Healthcare	941	1070	2011
2	Hotel →Healthcare	929	1082	2011

Classification Accuracy

To validate annotation quality, a Bi-LSTM classifier (128 units, 200 epochs) was trained on the labelled dataset. As shown in Table X, accuracy increased with dataset size and training ratio, reaching 74.60 % on a 90–10 split. Lower accuracy on smaller subsets reflects limited training data rather than annotation noise.

TABLE X

ACCURACY ON VARYING TRAINING-TESTING SPLITS

	ACCURACY (%)		
SIZE OF DATASET	70–30	80–20	90–10
500	62.00	61.00	65.00
1000	62.00	64.50	68.00
2000	68.14	69.76	74.60

Validity Analysis

Generalizability was evaluated by using annotated healthcare dataset as a source to label benchmark Product, Restaurant, and Hotel datasets. The resulting accuracies were comparable to standard source-target pairs (Fig. 9), confirming reliable annotation quality. The balanced fake-real distribution aligns with prior studies on online review manipulation [29]. This annotated healthcare dataset will be released publicly for future research.

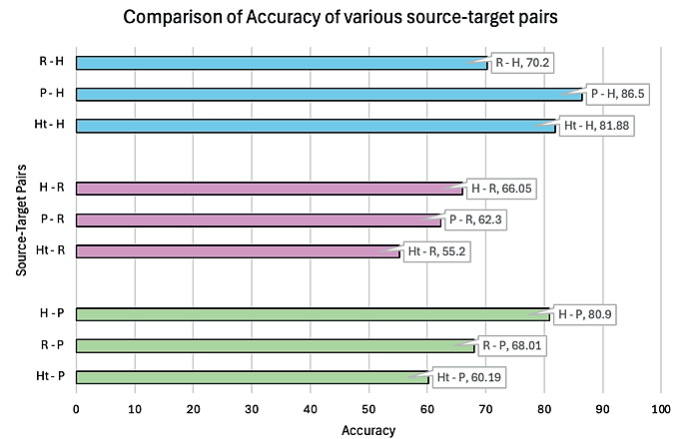


Fig. 9. Comparison of the accuracy percentages of various source-target pairs.

Expert Validation of Healthcare Dataset Labels: To assess the reliability of the automatically generated labels, a partial expert validation study was conducted on a random subset of 150 healthcare reviews. Two domain experts with prior experience in text annotation and healthcare-related content analysis independently reviewed the samples without access to the CORAL-generated labels. The agreement between CORAL-generated labels and expert annotations was found to be 82 %, indicating reasonably high level of labelling accuracy. However, a limitation of this approach is that some degree of labelling noise may persist due to domain shift and the absence of ground-truth labels in the target domain. Nevertheless, the validation results suggest that the generated labels are sufficiently reliable for supplemental evaluation.

VII. CONCLUSION

This study addressed the growing challenge of cross-domain fake review detection caused by limited labelled data and inherent class imbalance in real-world e-commerce reviews. To this end, a unified retrieval-augmented framework, CORAL, was proposed to enable automated, zero-shot fake review detection across domains without human intervention.

Extensive evaluation on widely used benchmark datasets demonstrated that CORAL achieved substantial improvements (average accuracy by 21.74 % and F1-score by 24.25 %) over existing baselines and prior research, while also reducing detection time. Beyond benchmark evaluation, CORAL was successfully applied to annotate a novel healthcare review dataset comprising 2011 reviews, directly addressing labelled data scarcity in socially critical domains.

Collectively, these results demonstrated CORAL's robustness, scalability, and practical suitability for real-world deployment. By enabling trustworthy, efficient, and domain-agnostic review moderation, CORAL supports informed consumer decision-making and strengthens platform credibility in competitive e-commerce environments.

REFERENCES

- [1] S. Paget, "Local consumer review survey 2025," Jan. 2025. [Online]. Available: <https://www.brightlocal.com/research/local-consumer-review-survey/>. Accessed on: November 2025.
- [2] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13079–13097, Jan. 2021. <https://doi.org/10.1007/s11042-020-10299-5>
- [3] R. Loconte, C. Battaglini, S. Maldera, P. Pietrini, G. Sartori, N. Navarin, and M. Monaro, "Detecting deception through linguistic cues: From reality monitoring to natural language processing," *Journal of Language and Social Psychology*, vol. 44, no. 3–4, pp. 523–552, 2025. <https://doi.org/10.1177/0261927X251316883>
- [4] N. Lebernegg, E. Jakob-Moritz, P. Tolochko, and H. Boomgaarden, "Do you speak disinformation? Computational detection of deceptive news-like content using linguistic and stylistic features," *Digital Journalism*, vol. 13, no. 8, pp. 1373–1398, 2025. <https://doi.org/10.1080/21670811.2024.2305792>
- [5] J. Hou, Z. Tan, S. Zhang, Q. Hu, and P. Wang, "Detecting fake review intentions in the review context: A multimodal deep learning approach," *Electronic Commerce Research and Applications*, vol. 70, Mar.–Apr. 2025, Art. no. 101485. <https://doi.org/10.1016/j.elerap.2025.101485>
- [6] J. M. Kim, K. K.-c. Park, M. Mariani, and S. F. Wamba, "Investigating reviewers' intentions to post fake vs. authentic reviews based on behavioral linguistic features," *Technological forecasting and social change*, vol. 198, Jan. 2024, Art. no. 122971. <https://doi.org/10.1016/j.techfore.2023.122971>
- [7] R. Gupta, V. Jindal and I. Kashyap, "Recent state-of-the-art of fake review detection: A comprehensive review," *The Knowledge Engineering Review*, vol. 39, Nov. 2024, Art. no. e8. <https://doi.org/10.1017/S0269888924000067>
- [8] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA, June 2014. <https://doi.org/10.3115/v1/P14-1147>
- [9] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: An empirical study," *Information Sciences*, vol. 385–386, pp. 213–224, Apr. 2017. <https://doi.org/10.1016/j.ins.2017.01.015>
- [10] L. C. Cagnina and P. Rosso, "Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 25, no. Suppl. 2, pp. 151–174, 2017. <https://doi.org/10.1142/S0218488517400165>
- [11] J. Sánchez-Junquera, L. Villaseñor-Pineda, M. Montes-y-Gómez, P. Rosso, and E. Stamatatos, "Masking domain-specific information for cross-domain deception detection," *Pattern Recognition Letters*, vol. 135, pp. 122–130, July 2020. <https://doi.org/10.1016/j.patrec.2020.04.020>
- [12] G. Ren, H. Wang, and Y. Yang, "Cross-domain fake review detection based on deep learning multilevel generic features extraction fusion," *Informatica*, vol. 49, no. 18, Apr. 2025. <https://doi.org/10.31449/inf.v49i18.7071>
- [13] R. Gupta, V. Jindal, and I. Kashyap, "Cross-domain fake review detection via orthogonal counterfactual representations," *IEEE Access*, vol. 13, pp. 119004–119022, July 2025. <https://doi.org/10.1109/ACCESS.2025.3584731>
- [14] K. Wang and X. Wan, "Counterfactual representation augmentation for cross-domain sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1979–1990, Mar. 2022. <https://doi.org/10.1109/TAFFC.2022.3158843>
- [15] J. Wu, Q. Liu, W. Xu, and S. Wu, "Bias mitigation for evidence-aware fake news detection by causal intervention," in *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, July 2022, pp. 2308–2313. <https://doi.org/10.1145/3477495.3531850>
- [16] Y. Tang, W. Wan, H. Chen, and M. Matsushita, "Zero-shot recognition of test tube types by automatically collecting and labeling RGB data," *IEEE Robotics and Automation Letters*, vol. 10, no. 8, pp. 8276–8283, July 2025. <https://doi.org/10.1109/LRA.2025.3585759>
- [17] M. G. Atigh, S. Nargang, M. Keller-Ressel, and P. Mettes, "SimZSL: Zero-shot learning beyond a pre-defined semantic embedding space," *International Journal of Computer Vision*, vol. 133, pp. 5161–5177, May 2025. <https://doi.org/10.1007/s11263-025-02422-6>
- [18] R. Zhang, Y.-S. Wang, and Y. Yang, "Generation-driven contrastive self-training for zero-shot text classification with instruction-following LLM," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, St. Julian's, Malta, Mar. 2024, pp. 659–673. <https://doi.org/10.18653/v1/2024-eacl-long.39>
- [19] R. A. Duma, Z. Niu, A. Nyamawe, J. Tchaye-Kondi, J. Chambua, and A. A. Yusuf, "DHMFRD – TER: a deep hybrid model for fake review detection incorporating review texts, emotions, and ratings," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4533–4549, May 2023. <https://doi.org/10.1007/s11042-023-15193-4>
- [20] J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, Jan. 2022, Art. no. 102771. <https://doi.org/10.1016/j.jretconser.2021.102771>
- [21] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, 2015, pp. 985–994. <https://doi.org/10.1145/2783258.2783370>
- [22] B. Ni, Z. Liu, L. Wang, Y. Lei, Y. Zhao, X. Cheng, Q. Zeng, L. Dong, Y. Xia, K. Kenthpadi, and R. Rossi, "Towards trustworthy retrieval augmented generation for large language models: A survey," *arXiv:2502.06872*, Feb. 2025. <https://doi.org/10.48550/arXiv.2502.06872>
- [23] C.-S. Wei, P.-Y. Hsu, C.-W. Huang, M.-S. Cheng, and G. F. Prassida, "Devising a cross-domain model to detect fake review comments," in *International Conference on Computational Collective Intelligence*, Nov. 2020, pp. 714–725. https://doi.org/10.1007/978-3-030-63119-2_58
- [24] Y. C. Chen, L. J. Chen, P.-Y. Hsu, M.-S. Cheng, and Y.-C. Chen, "Identify deceptive reviews in cross-domain content with BERT," in *International Conference on Pattern Recognition and Artificial Intelligence*, Feb. 2025, pp. 31–46. https://doi.org/10.1007/978-981-97-8705-0_3
- [25] GeeksforGeeks, "Perplexity for LLM evaluation," July 2025. [Online]. Available: <https://www.geeksforgeeks.org/nlp/perplexity-for-llm-evaluation/>. Accessed on: 22 september 2025.
- [26] S. Liu, K. Halder, Z. Qi, W. Xiao, N. Pappas, P. M. Htut, N. A. John, Y. Benajiba, and D. Roth, "Towards long context hallucination detection," *arXiv:2504.19457*, Apr. 2025. <https://doi.org/10.48550/arXiv.2504.19457>
- [27] P. Jonaitis, "Instant data scraper," Jan. 2025. [Online]. Available: <https://chromewebstore.google.com/detail/instant-data-scraper/foakhiiedipchpaobibbnahnkdoiiah?pli=1>. Accessed on: June 2025.

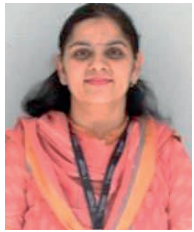
- [28] S. Panda and S. Levitan, "Deception detection within and across domains: Identifying and understanding the performance gap," *ACM Journal of Data and Information Quality*, vol. 15, no. 1, pp. 1–27, 2022.
<https://doi.org/10.1145/3561413>
- [29] Q. Wang, W. Zhang, J. Li, Z. Ma, and J. Chen, "Benefits or harms? The effect of online review manipulation on sales," *Electronic Commerce Research and Applications*, vol. 57, Jan.–Feb. 2023, Art. no. 101224.
<https://doi.org/10.1016/j.elerap.2022.101224>



Richa Gupta is an Assistant Professor with more than 14 years of experience at the Department of Computer Science, Keshav Mahavidyalaya, University of Delhi. She did her M.Sc. in Computer Science from University of Delhi in 2008, and B.Sc. in Computer Science from the University of Delhi in 2006. She mainly works in the area of artificial intelligence and natural language processing. She is currently pursuing her PhD from Manav Rachna International Institute of Research and Studies. Her areas of interest include machine learning, deep learning, generative AI, cybersecurity, and cloud computing.

E-mail: richa.gupta@keshav.du.ac.in

ORCID iD: <https://orcid.org/0000-0001-6094-4108>



Indu Kashyap has 15 years of teaching, administration, and research experience. She is a Professor at Manav Rachna International Institute of Research and Studies. She has a PhD from Chaudhary Charan Singh University, Meerut. She has over 40 publications in reputed journals and conferences, including Elsevier, Springer, and Taylor & Francis. Her research areas include wireless networks, machine learning, data analytics, and recommender systems.

E-mail: indu.set@mriu.edu.in

ORCID iD: <https://orcid.org/0000-0003-1884-0828>



Vinita Jindal is a professor with more than 24 years of experience at the Department of Computer Science, Keshav Mahavidyalaya, University of Delhi. Before joining the Keshav Mahavidyalaya, she worked as a Sr. Faculty in the PCTI Ltd. from July 1999 to July 2001. She did her Doctorate in Computer Science from the University of Delhi in 2018. She did her M.Phil. in Computer Science from Madurai Kamaraj University in 2007, MCA from IGNOU in 2000 and B.Sc. in Mathematics from the University of Delhi in 1997. She mainly works in the area of artificial intelligence and computer networks. Her areas of interest include neural networks, machine learning, cybersecurity, intrusion detection systems, dark web, deep learning, recommender systems and vehicular ad-hoc networks. She has published more than 40 research articles in good impact journals, book chapters and reputed conference proceedings.

E-mail: vjindal@keshav.du.ac.in

ORCID iD: <https://orcid.org/0000-0002-0481-4840>

APPENDIX A

This appendix provides the high-level pseudocode of CORAL, summarising its four phases and the auxiliary module.

Pseudocode for CORAL:

A Unified Counterfactual Transfer Learning Framework

Input: Source Dataset $D_S = \{(r_i^S, l_i^S)\}_{i=1}^{N_S}$, Target Dataset $D_T = \{(r_j^T)\}_{j=1}^{N_T}$,

$Params = \{\delta, \tau, k, imbalance_flag, embedding_model, retriever_model\}$

Output: Predicted labels for target dataset L_T

```
%----- Auxiliary Module -----%
1 if imbalance_flag == TRUE then
2    $D_S \leftarrow augment(D_S)$ 
3 end
%----- Phase I: Unified Masking -----%
4 Function unified_masking( $D_S, D_T, \delta, \tau$ ):
5   compute domain-affinity scores  $\rho(w, D_S)$  and  $\rho(w, D_T)$  for vocabulary
6   compute heuristic score  $M_1(w, D_S, D_T)$ 
7   train or use domain classifier to obtain attention-based score  $M_2(w, D_S)$ 
8   for each token  $w$  in each source review  $r_i^S$  do
9      $M(w) \leftarrow \delta \cdot M_1(w) + (1 - \delta) \cdot M_2(w)$ 
10    if  $M(w) \geq \tau$  then replace  $w$  with [MASK]
11  end
12 return masked_reviews
13 end
%----- Phase II: Counterfactual Generation -----%
14 % Build index on target domain for RAG
15 Index  $T \leftarrow Build\_Index(D_T, embedding\_model)$ 
16 Function Generate_counterfactual(masked_reviews, Index  $T, k, prompt$ ):
17    $Q \leftarrow Retrieve\_topK(masked\_reviews, Index\_T, k)$  %top-k relevant target chunks
18   prompt  $\leftarrow format\_prompt(prompt, masked\_reviews, Q)$ 
19    $C \leftarrow Gemini\_generate(prompt, masked\_reviews, Q)$ 
20   return  $C$  %counterfactual review
21 end
22  $C_{set} \leftarrow \{\}$ 
23 for each  $R_i$  in masked_reviews do
24    $C_i \leftarrow Generate\_counterfactual(R_i, Index\_T, k, prompt)$ 
25   store label  $l_i^S$  with  $C_i$ 
26    $C_{set} \leftarrow C_{set} \cup \{C_i, l_i^S\}$ 
27 end
%----- Phase III: Orthogonal alignment -----%
29 % obtain embeddings for counterfactuals and representative target samples
30  $X \leftarrow embed(C_i \text{ from } C_{set}, embedding\_model)$ 
31  $Y \leftarrow sample\_target\_embeddings(D_T, embedding\_model)$ 
32 %Perform OPA and replace counterfactuals with aligned versions
33  $C_{alignset} \leftarrow \{(X_{aligned_i}, l_i^S)\}$ 
%----- Phase IV: Classification -----%
34 % Train classifier on aligned counterfactuals
35 model  $\leftarrow Train_{BiLSTM}(C_{alignset})$ 
36 % Predict labels for target reviews
37 for each  $r_j^T$  in  $D_T$  do
38    $\hat{Y}_j \leftarrow predict(model, r_j^T)$ 
39    $L_T.add(\hat{Y}_j)$ 
40 end
41 return  $L_T$ 
```

APPENDIX B

Qualitative error analysis using representative cases

To enhance interpretability and transparency, this appendix presents representative case studies illustrating CORAL's classification behaviour across tasks. For each case, the original review text, original label, predicted label, error type, and a likely cause of misclassification are reported in Table XI.

These examples demonstrate how CORAL's predictions are grounded in retrieved contextual evidence, while also highlighting

typical failure modes under domain shift. False negatives often arise when domain-specific service cues or exaggerated praise are insufficiently captured during adaptation, whereas false positives occur for concise or generic positive reviews that resemble promotional templates. Overall, the cases provide qualitative insight into the strengths and limitations of CORAL, complementing the quantitative evaluation reported in Section V, and supporting explainability requirements for trustworthy AI in e-commerce review moderation.

TABLE XI
REPRESENTATIVE ERROR CASES ACROSS TASKS

TASK	REVIEW TEXT	TRUE	PREDICTED	ERROR TYPE	LIKELY CAUSE
Product→Restaurant	The food arrived late, but overall not too bad.	Fake	Genuine	FN	Domain-specific service cues missed
Product→Restaurant	Authentic Italian pizza with crispy crust, will visit again!	Genuine	Fake	FP	Promotional style misclassified
Product→Hotel	The check-in was smooth and the view was breathtaking.	Fake	Genuine	FN	Unseen hotel terminology
Product→Hotel	Room was spacious and comfortable for our stay.	Genuine	Fake	FP	Positive but simple descriptions misclassified
Hotel→Product	Battery lasts only two hours, totally useless.	Fake	Genuine	FN	Very negative review dismissed due to domain mismatch
Hotel→Product	This phone cover fits perfectly and protects from scratches.	Genuine	Fake	FP	Short direct praise confused as fake
Hotel→Restaurant	Chef personally came to greet us, unforgettable evening.	Fake	Genuine	FN	Flowery praise style overlooked
Hotel→Restaurant	Tried their sushi platter, fresh and flavorful.	Genuine	Fake	FP	Concise positive review confused as fake
Restaurant→Product	Headphones stopped working in two weeks, worst purchase.	Fake	Genuine	FN	Harsh negativity seen as genuine complaint
Restaurant→Product	Lightweight mouse, fits hand comfortably.	Genuine	Fake	FP	Template-like brevity confused as fake
Restaurant→Hotel	Five stars for the friendly staff and stunning lobby!	Fake	Genuine	FN	Exaggerated compliments overlooked
Restaurant→Hotel	Stayed three nights, clean room and quiet atmosphere.	Genuine	Fake	FP	Concise praise resembled boilerplate fake