Debiasing Logical Fallacy Detection for Real-World Robustness via Counterfactually Augmented Data

Navyansh Singh
Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur
Raipur, Chhattisgarh, India
navyansh24102@iiitnr.edu.in

Abstract

Logical fallacy detection models frequently over-flag valid reasoning due to reliance on surface-level spurious correlations. We introduce 703 LLM-generated Counterfactually Augmented Data (CAD) pairs—minimally differentiated valid and fallacious arguments—to debias models through targeted augmentation. Fine-tuning DeBERTa-v3-large on CoCoLoFa augmented with these pairs yields marginal in-distribution improvement (+0.4% F1) but substantial out-of-distribution robustness: 58% relative reduction in false positive rate (64% $\rightarrow$ 26.7%) on a 300-sample Reddit-sourced evaluation set. While recent LLMs (Llama-3.1-8B, Llama-3.3-70B) achieve high performance under optimized prompts (F1 90–94%), they degrade severely under simple human-like prompts (F1 63–72%, FPR 54–74%). Our lightweight, prompt-invariant approach achieves competitive robustness (F1 85.9%, FPR 26.7%) across all prompting regimes without prompt engineering, making it stable for production deployment with unpredictable user input. The dataset and model are publicly released.

Keywords: logical fallacy detection, counterfactual augmentation, debiasing, real-world robustness, prompt-independence, LLM evaluation

1 Introduction

Logical fallacy detection aims to identify flawed reasoning in text, with applications in misinformation mitigation, debate analysis, and content moderation. Recent benchmarks such as CoCoLoFa (Yeh et al., 2024) report strong in-distribution performance (F1 $\approx$ 0.86 for fine-tuned BERT variants), yet these models often exhibit high false positive rates on logically valid arguments in real-world settings. Manual inspection of CoCoLoFa revealed consistent over-flagging of non-fallacious statements—particularly those with authoritative tone, causal structure, or common reasoning markers (e.g., "therefore", "doctor")—suggesting over-reliance on surface patterns rather than genuine logical understanding.

This over-flagging limits practical deployment in automated pipelines such as content moderation systems and debate analysis tools, where text is diverse, noisy, and context-dependent, and where flagging valid arguments as fallacious carries real costs. While large language models (LLMs) show promise in zero-shot fallacy detection under optimized conditions, their performance remains highly sensitive to prompt formulation and lacks the efficiency required for high-throughput applications. This prompt sensitivity is a fundamental liability in production environments where user input is unpredictable and cannot be pre-engineered with multi-paragraph instructions. Neurosymbolic and hybrid approaches (Lalwani et al., 2025; Kutepova and Khatib, 2025) address logical structure but introduce complexity and computational overhead.

We hypothesize that targeted counterfactual augmentation (Kaushik et al., 2020)—creating minimally differentiated valid-fallacious pairs—can refine model decision boundaries to prioritize logical cues over spurious correlations, improving robustness without degrading in-distribution performance. We test this by generating and manually verifying 703 such pairs aligned with CoCoLoFa's eight fallacy types, then fine-tuning DeBERTa-v3-large on augmented data.

We pursue three main objectives: (1) the release of a publicly available dataset containing 703 high-quality Counterfactually Augmented Data (CAD) pairs designed for fallacy debiasing; (2) the empirical demonstration of substantial gains in out-of-distribution robustness, including a 58% relative reduction in false positive rate on a balanced 300-sample manual evaluation set; and (3) the provision of evidence that a prompt-invariant, lightweight model serves as a reliable alternative to prompt-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 363–374

sensitive large language models in real-world settings with unpredictable user input.

2 Related Work

2.1 Logical Fallacy Detection

Logical fallacy detection has gained increasing attention as a subtask of argument mining and reasoning evaluation. Early datasets such as those focused on ad hominem fallacies (Habernal et al., 2018) and the LOGIC dataset (Jin et al., 2022) emphasized formal or synthetic arguments but were limited in size and real-world domain coverage.

The most recent comprehensive benchmark is CoCoLoFa (Yeh et al., 2024), a dataset of 7,706 real-world news comments labeled with eight common fallacy types via LLM-assisted crowdsourcing. Fine-tuned transformer models (e.g., BERT, RoBERTa, DeBERTa) achieve strong in-distribution F1 scores of $\approx$0.86. However, analyses within the paper and related work highlight persistent brittleness: models frequently over-rely on surface cues (e.g., keywords like "therefore", authoritative tone, or causal phrasing), leading to high false positive rates on valid reasoning. Complementary datasets like Missci (Glockner et al., 2024) focus on fallacies in scientific claims, addressing implicit fallacious reasoning by grounding in misrepresented publications, while SLURG (Blanco et al., 2025) explores synthetic generation of fallacious forum-style comments to augment informal discourse data.

Gap in Prior Work: While these benchmarks report strong in-distribution performance, they do not adequately address the brittleness problem—models' over-reliance on superficial patterns limits real-world deployment. Our work directly targets this robustness gap through counterfactual debiasing.

2.2 Counterfactual Augmentation and Debiasing

Counterfactual data augmentation (Kaushik et al., 2020) creates minimally-differentiated example pairs to force models to learn causal distinctions rather than spurious correlations. Originally applied to sentiment analysis, Kaushik et al. (2020) generated positive/negative review pairs differing only in sentiment-critical words (e.g., "The food was *delicious*" vs. "The food was *terrible*"). This approach compels models to focus on truly predictive features rather than dataset-specific shortcuts.

We adapt this methodology to logical fallacy detection by creating valid/fallacious argument pairs that differ *only* in logical soundness while preserving topic, structure, tone, and most wording. For example:

- **Valid:** "The plant died because it received no sunlight for two weeks—plants require photosynthesis to survive."
- **Fallacious:** "The plant died shortly after I played rock music, so the music must have killed it." [post hoc fallacy]

Both sentences contain causal markers and similar structure, but only the logical relationship differs. Training on such pairs prevents models from achieving low loss through surface-level shortcuts—they must learn to distinguish sound versus flawed reasoning.

While prior CDA methods hold surface features constant across semantically related domains—sentiment polarity in reviews (Kaushik et al., 2020), factual stance in political text (Sermsri and Panboonyuen, 2025), or bias attributes via semantic triples (Jin et al., 2025)—our pairs introduce a qualitatively different constraint: logical validity must flip while *all* surface features (topic, syntactic structure, length, tone, and most wording) are held fixed. This is a stricter requirement than sentiment or stance debiasing, because logical validity is not recoverable from any single lexical or syntactic cue—it depends on the inferential relationship between premises and conclusion. A model cannot shortcut through word choice, connectives, or tone; it must learn to evaluate the logical link itself. This design choice is what distinguishes our CAD pairs from related approaches such as MisSynth (Poliakov and Shvai, 2025), which employs RAG-based generation to create fallacious scientific claims but produces loosely related examples rather than minimally differentiated pairs that precisely isolate logical distinctions.

Contrastive learning methods inspired by SimCLR (Chen et al., 2020), adapted to text (Gao et al., 2021), and hard-negative mining (Karpukhin et al., 2020) have boosted performance in entailment, question answering, and reasoning tasks. In argumentation, contrastive techniques appear in claim verification (Thorne et al., 2018) and natural language inference (Nie et al., 2020), though applications to fallacy detection remain limited. Two early works on shortcut learning in argument

comprehension are also directly relevant: Niven and Kao (2019) demonstrated that NLI models exploit dataset-specific lexical artefacts rather than genuine argument understanding, and Kavumba et al. (2019) showed analogous shortcut behaviour in plausible alternative selection—both motivating the need for contrastive debiasing in reasoning tasks.

Our Contribution: We generate 703 tightly matched valid-fallacious CAD pairs specifically designed for fallacy detection, extending Kaushik et al. (2020)'s approach to logical reasoning. Unlike prior CDA work, which targets surface-level attributes such as sentiment or stance, our pairs enforce minimal differentiation at the level of logical structure—the most fine-grained and task-critical axis of variation for fallacy detection.

2.3 Large Language Models for Fallacy Detection

Zero-shot and few-shot prompting of large language models shows promise for fallacy detection, with evaluations of models like GPT-4 demonstrating high F1 under optimized prompts. Nevertheless, performance is highly sensitive to prompt wording, example selection, and model scale (Webson and Pavlick, 2022; Kojima et al., 2023). Studies report elevated false positive rates under simple, human-like prompts (Payandeh et al., 2023)—a critical weakness for production environments where user input is unpredictable and cannot be pre-engineered with multi-paragraph instructions.

Recent explorations further examine LLM vulnerability to logical fallacies in scientific reasoning, addressing fallacies via domain-specific prompting or hybrid approaches (Poliakov and Shvai, 2025; Jeong et al., 2025; Sourati et al., 2023). Additional critiques highlight fallacies in equating LLMs with human reasoning (Lin, 2025) and dissociating language from thought (Mahowald et al., 2024).

Our Approach: Rather than relying on prompt engineering, we develop a lightweight, prompt-invariant fine-tuned model that maintains stable performance regardless of input phrasing, making it more reliable for real-world deployment where prompts cannot be controlled.

2.4 Real-World and Out-of-Distribution Evaluation

A growing literature critiques over-reliance on in-distribution benchmarks and advocates ecologically valid evaluation—testing on real-world, naturalistic data that reflects actual deployment conditions (Bowman and Dahl, 2021; Ribeiro et al., 2020). In argumentation and reasoning, manual out-of-distribution sets (e.g., adversarial examples or user-generated discourse from forums) reveal robustness gaps that standard benchmarks miss. Related work includes ATOM (Chen et al., 2021) for robust OOD detection via outlier mining and robustness evaluation frameworks for argument mining (Sofi et al., 2022), which develop methods to assess and improve model performance beyond aggregate metrics.

Our Evaluation: Our primary evaluation uses a manually curated, balanced set of 300 real-world examples (150 valid + 150 fallacious), predominantly from Reddit discussions, aligning with this emphasis on noisy, diverse, user-generated text over controlled benchmarks. Approximately 60% of valid examples are "hard negatives"—valid arguments containing superficial triggers (e.g., "experts say," "therefore") that mimic common fallacy patterns, specifically designed to stress-test models' reliance on shortcuts.

2.5 Positioning of This Work

This work bridges these research threads by: (1) applying targeted counterfactual augmentation (inspired by Kaushik et al. (2020)) specifically to debias logical fallacy detectors; (2) demonstrating substantial robustness gains on challenging real-world data where existing models fail; and (3) providing evidence that lightweight, prompt-invariant fine-tuned models serve as reliable alternatives to prompt-sensitive large language models in practical deployment scenarios.

3 Methodology

Our approach consists of three components designed to debias logical fallacy detection models. First, we generate 703 CAD pairs (Section 3.1) that isolate logical distinctions from surface-level patterns through minimal differentiation. Second, we fine-tune DeBERTa-v3-large on CoCoLoFa augmented with these pairs (Section 3.2), enabling the model to learn genuine logical reasoning rather than dataset-specific shortcuts. Third, we evaluate on a challenging 300-sample real-world manual set (Section 3.3) to measure robustness beyond in-distribution benchmarks. This design tests our central hypothesis: minimally-differentiated valid/fallacious pairs refine model decision bound-

aries to prioritize logical validity over spurious correlations, yielding models robust to real-world discourse.

3.1 Counterfactually Augmented Data Generation

To encourage the model to focus on genuine logical distinctions rather than superficial lexical or syntactic patterns, we generate 703 high-quality CAD pairs using GPT-4 (gpt-4-0613, queried November–December 2024, knowledge cutoff September 2021). Each pair consists of:

- A valid argument grounded in sound reasoning.
- A minimally differentiated fallacious counterpart that introduces exactly one logical fallacy while preserving the topic, syntactic structure, length, tone, and most of the wording.

This "minimal differentiation" design creates hard negatives that force the model to learn subtle reasoning cues rather than relying on surface-level heuristics.

Mechanistic Rationale: Counterfactual pairs work by creating training examples where surface features are held constant while logical validity flips. Consider the example pair:

- **Valid:** "He hasn't eaten in 12 hours, so he is probably hungry."
- **Fallacious:** "He hasn't eaten in 12 hours, so he must be a terrible guitar player." [non sequitur]

Both sentences share identical structure ("He hasn't eaten in 12 hours, so..."), similar length, and neutral tone. A model relying on shortcuts—such as flagging arguments containing "so" or causal phrasing as fallacious—would misclassify both identically. However, their labels differ based purely on *logical structure*: the first establishes a plausible causal relationship (hunger follows from lack of food), while the second makes an entirely unwarranted inference with no logical connection between premises and conclusion. By training on such pairs, the model cannot minimize loss through surface-level pattern matching—it must learn to distinguish sound causal reasoning from flawed inference. This forces the decision boundary to align with logical validity rather than lexical or syntactic cues prevalent in single-source datasets like CoCoLoFa.

Pairs were generated in batches aligned with the eight fallacy types present in CoCoLoFa (e.g., ad hominem, strawman, false cause, appeal to emotion, etc.). The generation prompt template was as follows:

> *Generate 30 pairs. Sentence A explains a proven causal mechanism. Sentence B assumes causation because one thing happened after another.*
>
> **Example**
>
> Valid: "The plant died because it was kept in a dark room and received no sunlight for two weeks."
>
> Fallacy: "The plant died shortly after I played a rock song in the room, so the music must have killed it."

After generation, all 703 pairs were manually reviewed and verified by the authors for label correctness and minimal differentiation quality. To further assess verification quality, a random subset of 50 pairs (100 individual arguments) was independently double-annotated by two raters for *binary valid/fallacious labels*, yielding raw agreement of 84.0% (16 disagreements out of 100 labels) and Cohen's $\kappa = 0.68$ (substantial agreement).

To prevent data leakage, we strictly ensured that the 300-sample manual evaluation set (described in Section 3.3) was never included in the GPT-4 prompts used for counterfactual generation.

This curated dataset is publicly released to support reproducibility and future research in fallacy debiasing.

Dataset Clarity: Two Separate Datasets

- **703 CAD Pairs:** Synthetic, GPT-4 generated, manually verified by the authors. Used for *training augmentation* only. Never exposed to test evaluation.
- **300 Manual Evaluation Set:** Real-world Reddit/forum arguments, human-curated. Used *exclusively* for out-of-distribution testing (Section 3.3). Zero overlap with training data.

3.2 Model Architecture and Training

DeBERTa-v3-large (He et al., 2023) is selected as our backbone due to its disentangled attention mechanism. Unlike standard BERT, which mixes content and position signals, DeBERTa represents them as separate vectors.

We hypothesize this disentangled representation is beneficial for fallacy detection because logical reasoning depends critically on relational structure: the validity of "A therefore B" depends on whether A *actually supports* B, not merely their textual proximity. By separating content (semantic meaning of words) from position (syntactic structure and connectives), DeBERTa can better model these compositional logical dependencies. Prior work shows DeBERTa outperforms RoBERTa on NLI tasks by $\sim$2% F1 (He et al., 2023), consistent with our observations on CoCoLoFa.

The model is fine-tuned for binary classification (valid vs. fallacious) on:

- CoCoLoFa training set (primary realistic data)
- Augmented with $3\times$ oversampled CAD pairs (2,109 additional examples)

$3\times$ oversampling is selected after comparing $0\times$, $1\times$, and $3\times$ variants (results in Table 2). While $1\times$ oversampling achieved perfect recall (100%), it suffered from a higher false positive rate (36% FPR) as the model became overly conservative in flagging arguments. $3\times$ strikes the optimal balance: substantial robustness gains (26.7% FPR, Table 2) without in-distribution degradation (86.6% F1, only +0.4% from baseline; Table 1). We did not explore higher oversampling factors (e.g., $5\times$) to avoid potential overfitting to synthetic data patterns.

Training on the augmented counterfactual data ensures that the model's strong performance on real-world Reddit evaluation (Section 4) results from learning general logical reasoning patterns rather than memorizing lexical shortcuts, as the synthetic training pairs and real-world test examples share no vocabulary overlap beyond common function words. Training hyperparameters are provided in Appendix A.

We compare three model variants:

- **cocopure**: Fine-tuned only on CoCoLoFa (baseline)
- **coco_1x**: Fine-tuned on CoCoLoFa + $1\times$ CAD pairs (no oversampling)
- **coco_pair**: Fine-tuned on CoCoLoFa + $3\times$ oversampled CAD pairs (our proposed model)

All models are trained with the same hyperparameters and random seed for fair comparison.

Practical Advantage: Unlike LLM-based approaches that require carefully engineered prompts (Section 4.3), our fine-tuned models are prompt-invariant—they operate on raw text input without any hand-crafted instructions or in-context examples at inference time. This makes them robust to the unpredictable phrasing typical of real-world user queries, a critical requirement for production deployment where users may phrase requests in diverse, uncontrolled ways.

3.3 Evaluation Protocol

We prioritize ecologically valid evaluation—testing on real-world, naturalistic data that reflects actual deployment conditions—over in-distribution benchmark performance. Concretely, we distinguish two evaluation settings: *in-distribution*, using the held-out CoCoLoFa test set (same domain and collection procedure as training data), and *out-of-distribution* (OOD), using the 300-sample Reddit-sourced set described below (different domain, source, and noise characteristics).

Our primary OOD evaluation uses a manually curated, balanced set of 300 real-world examples (150 valid + 150 fallacious), predominantly from Reddit discussions (2023–2025 threads) in subreddits such as r/changemyview, r/askphilosophy, r/Ask_Politics, r/DebateReligion, r/DebateAVegan, r/logic, r/philosophy, r/The10thDentist, r/climatechange, r/privacy, and others. Examples were collected by the authors via targeted keyword searches (e.g., "ad hominem", "post hoc", "is this a fallacy?") combined with domain terms, drawing from $\sim$300–400 high-engagement posts. The authors selected and lightly paraphrased the final examples to preserve messy, natural tone (imperfect phrasing, emotions, biases) while ensuring label clarity; paraphrasing was performed prior to and independently of the annotation step.

Fallacious cases show clear logical errors; valid examples include approximately 60% "hard negatives"—valid arguments containing superficial triggers (e.g., "experts say", "therefore", "because") that mimic common fallacy patterns in CoCoLoFa. This design specifically stress-tests whether models rely on shortcuts versus genuine logical understanding. Approximately 30 examples

per category were supplemented from educational sources (e.g., Grammarly, BBC) for variety. All were human-reviewed for label accuracy and diversity across domains, lengths, and fallacy subtypes.

To validate the test set, two independent annotators labeled a subset of 60 samples. We report a raw agreement of 88.3% and Cohen's $\kappa = 0.77$, indicating substantial agreement. Disagreements occurred primarily on nuanced "gray area" arguments (discussed in Section 4.5).

Metrics include:

- Overall accuracy and F1-score
- Precision, recall
- **False Positive Rate (FPR)** on valid examples (core robustness metric—the proportion of valid arguments wrongly flagged as fallacious)
- False Negative Rate (FNR) on fallacious examples

For completeness, we also report in-distribution F1 on the CoCoLoFa test set (Table 1).

Additionally, we evaluate recent open-weight LLMs (Llama-3.1-8B and Llama-3.3-70B) under three prompting regimes:

- Zero-shot with optimized prompts
- Simple human-like prompt ("Does this argument have a logical fallacy? Answer yes or no:")
- Few-shot with 4 human-written examples (2 valid, 2 fallacious)

This multi-regime comparison highlights prompt sensitivity and demonstrates the advantages of our lightweight, prompt-invariant approach across varying deployment scenarios.

Resources. All code, the 703-pair CAD dataset, and the fine-tuned model are publicly released.[1]

4 Experiments

4.1 Experimental Setup

We evaluate our models on both in-distribution benchmarks and the out-of-distribution manual test set described in Section 3.3.

[1]Dataset: https://huggingface.co/datasets/Navy0067/contrastive-pairs-for-logical-fallacy
Model: https://huggingface.co/Navy0067/Fallacy-detector-binary

In-distribution benchmarks Table 1 reports F1-score on the CoCoLoFa test set (Yeh et al., 2024). Our augmented model (**coco_pair**) achieves 86.6% F1, only +0.4% above the baseline—the expected outcome, since our goal is real-world robustness rather than benchmark optimization.

Model	F1	Acc	Prec	Rec
cocopure	86.2	86.4	85.1	87.3
coco_pair (ours)	**86.6**	**86.8**	**86.0**	**87.3**

Table 1: CoCoLoFa test set performance (single run, same random seed). All metrics reported as percentages. Marginal improvement confirms that augmentation preserves in-distribution quality while targeting real-world robustness.

Real-world manual evaluation set As detailed in Section 3.3, we use a balanced 300-sample out-of-distribution set (150 valid + 150 fallacious) from Reddit and other sources, with ∼60% hard negatives to stress-test robustness.

4.2 Results

We observe that the baseline's 64% FPR on our challenging set confirms the shortcut-learning hypothesis: the model associates reasoning markers with fallacies, failing when used correctly. Our counterfactual-augmented model achieves a 58% relative reduction in false positive rate (64.0% → 26.7%) while maintaining high recall on fallacious examples. While coco_1x achieved a perfect 100% recall, it suffered from a higher FPR (36.0%). Our final coco_pair model (3×) offers the best balance for real-world utility by significantly further reducing false alarms (26.7% FPR) with only a marginal trade-off in recall.

Model	Acc	F1	Prec	Rec	FPR	FNR
cocopure	60.0	67.7	56.8	84.0	64.0	16.0
coco_1x	82.0	84.7	73.5	100.0	36.0	0.0
coco_pair	**84.3**	**85.9**	**78.1**	**95.3**	**26.7**	**4.7**

Table 2: Overall performance on the 300-sample real-world manual evaluation set. All metrics reported as percentages.

Model	Acc (Valid)	False Pos	FPR
cocopure	36.0	96	64.0
coco_1x	64.0	54	36.0
coco_pair	**73.3**	**40**	**26.7**

Table 3: Performance on valid examples only (150 valid arguments). Accuracy and FPR reported as percentages.

Our model correctly identifies 73.3% of valid real-world arguments as valid, compared to only 36.0% for the baseline—a substantial improvement in robustness.

4.3 Comparison with Large Language Models

We observe that LLMs collapse under simple human prompts (FPR up to 74%), confirming that scale alone does not guarantee logical robustness. We hypothesize this reflects a safety-first bias—over-flagging assertive reasoning unless explicitly instructed otherwise—a fundamental liability where prompt engineering cannot be guaranteed. Notably, Llama-3.1-8B outperforms the larger Llama-3.3-70B under optimized prompts (F1 94.5 vs. 90.8), suggesting prompt sensitivity does not scale monotonically with model size. In contrast, our prompt-invariant model maintains stable decision boundaries and high prediction confidence (average softmax probability of 99.3% on correctly classified examples) across all conditions without any prompt engineering.

4.4 Error Analysis

We manually inspected the 40 false positives and 7 false negatives from our best model (**coco_pair**). The asymmetry in counts directly reflects the model's error rates: with a 26.7% FPR on 150 valid examples, 40 false positives are expected; with a 4.7% FNR on 150 fallacious examples, only approximately 7 false negatives arise naturally, making this a considerably rarer error type.

False Positives (valid arguments wrongly flagged as fallacious) The model remains somewhat oversensitive to surface cues resembling common fallacies:

- Credibility/conflict-of-interest critiques misclassified as ad hominem (e.g., valid funding bias concerns).
- Strong binary/excluded-middle reasoning flagged as false dichotomies.
- Evidence-backed slippery slope or induction arguments treated as unfounded.

False Negatives (fallacious arguments wrongly accepted as valid) The model is under-sensitive to subtler fallacies lacking strong surface cues:

- Pure appeal to emotion/pity (e.g., guilt-based manipulation).
- Tu quoque / identity-based ad hominem (e.g., dismissing based on speaker's group membership).
- Post hoc ergo propter hoc (causation from temporal sequence alone).

These patterns suggest that CAD pairs were highly effective at reducing superficial overfitting but left gaps in coverage for softer emotional appeals and abstract causal fallacies. Future work could extend the counterfactual dataset to target these subtypes.

4.5 Analysis of Human Ambiguity

Our inter-annotator agreement study (Cohen's $\kappa = 0.77$ on a subset of 60 samples) revealed that even for humans, some real-world arguments occupy a logical "gray area." For example, the statement: "Sally doesn't believe in God, so she must be an atheist" triggered disagreement: one rater identified a False Dichotomy (omitting agnosticism), while another viewed it as a valid definitional claim. Similarly, emotive appeals like "Think of the poor families being torn apart!" were contested between Valid Ethical Pathos and a fallacious Appeal to Pity. These instances suggest that the remaining 26.7% FPR in our model may target arguments that are inherently linguistically ambiguous, highlighting the inherent difficulty of the task even for human annotators.

5 Discussion

Our results demonstrate that targeted counterfactual augmentation significantly enhances the robustness of fallacy detection models. By introducing minimally differentiated valid-fallacious pairs, the model learns to prioritize genuine logical distinctions over superficial lexical and syntactic shortcuts prevalent in news-commentary datasets like CoCoLoFa. The 58% relative reduction in false positive rate (64.0% $\rightarrow$ 26.7%) on our challenging real-world evaluation set underscores the effectiveness of this approach in addressing over-flagging of valid reasoning.

Comparisons with large language models further highlight the advantages of task-specific debiasing. While Llama models excel under carefully engineered zero-shot prompts (F1 90–94%), they collapse under simple human-like prompts (FPR rising to 54–74%). This prompt sensitivity is a fundamental liability in production environments

Model	Prompt Regime	Acc	F1	Prec	Rec	FPR	FNR
Llama-3.1-8B	Zero-shot (optimized)	94.3	94.5	91.3	98.0	9.3	2.0
Llama-3.3-70B	Zero-shot (optimized)	90.0	90.8	84.1	98.7	18.7	1.3
Llama-3.3-70B	Simple human prompt	63.0	73.0	57.5	100.0	74.0	0.0
Llama-3.1-8B	Simple human prompt	72.3	78.1	64.6	98.7	54.0	1.3
Llama-3.3-70B	Few-shot (4 examples)	82.7	85.1	74.5	99.3	34.0	0.7
Llama-3.1-8B	Few-shot (4 examples)	89.0	90.0	82.7	98.7	20.7	1.3
coco_pair (ours)	Fine-tuned	**84.3**	**85.9**	**78.1**	**95.3**	**26.7**	**4.7**

Table 4: LLM performance across prompting regimes on the 300-sample set. All metrics reported as percentages.

where user input is unpredictable and cannot be pre-engineered with multi-paragraph instructions. In contrast, our prompt-invariant, lightweight model achieves competitive robustness (F1 85.9%, FPR 26.7%) without any prompt engineering or massive inference costs, making it more reliable for practical, high-throughput deployment.

6 Limitations

Despite these gains, several limitations remain.

First, the model still exhibits a residual false positive rate of 26.7% on valid arguments. Error analysis revealed over-sensitivity to certain surface cues (e.g., credibility critiques misclassified as ad hominem, strong binary framing flagged as false dichotomies). Conversely, subtler fallacies lacking strong lexical markers (e.g., pure appeal to emotion/pity, tu quoque, post hoc) are occasionally missed.

Second, while the 300-sample manual evaluation set is diverse and challenging, it remains relatively small compared to large-scale benchmarks, and is drawn from a single domain of user-generated Reddit content. Evaluation on additional clearly distinct domains—such as scientific debate platforms or formal parliamentary discourse—would provide stronger evidence for the generality of the proposed debiasing method. Human annotators themselves showed substantial but imperfect agreement (Cohen's $\kappa = 0.77$), with disagreements primarily on "gray area" arguments (e.g., distinguishing valid ethical pathos from fallacious appeal to pity), highlighting the inherent subjectivity in real-world fallacy detection.

Third, all experiments use a single random seed. Transformer-based models such as DeBERTa are known to exhibit variance across seeds (Dodge et al., 2020), and reporting mean performance with standard deviation over multiple runs would provide a more complete picture of reliability. We leave multi-seed evaluation and cross-architecture validation to future work.

Fourth, robustness gains are demonstrated only for DeBERTa-v3-large. Whether the benefits of CAD-based augmentation generalise to other architectures (e.g., RoBERTa, ELECTRA) or to different model sizes within the same family remains an open question. The disentangled attention mechanism of DeBERTa may contribute to its receptiveness to this form of augmentation, a hypothesis that warrants investigation.

Finally, the current binary classification setup does not distinguish between fallacy types. Extending to multi-class detection would enable finer-grained analysis and more targeted debiasing. Additionally, the counterfactual pairs were generated using GPT-4 and verified primarily by the authors; scaling to thousands of pairs with broader multi-annotator validation could further reduce bias and strengthen generalization.

7 Ethical Considerations & Broader Impact

This work aims to improve trustworthy reasoning detection for applications like misinformation mitigation and online moderation. By reducing false positives on valid arguments, our approach helps avoid over-censorship of legitimate discourse. However, misuse of fallacy detectors could amplify biases if deployed without careful monitoring.

All data used in the manual evaluation set is publicly available and anonymized. The counterfactual pairs are synthetic and do not contain real user information. We release all resources openly to encourage responsible research and deployment.

8 Conclusion & Future Work

We achieve a 58% relative reduction in false positives (64% $\rightarrow$ 26.7% FPR) on real-world arguments while maintaining strong fallacy detection. Even large LLMs collapse under natural prompts (FPR 54–74%), suffering 'safety-first' over-flagging bias. Our task-specific, prompt-invariant model competes with few-shot 70B models while being orders

of magnitude cheaper and faster.

This suggests targeted data augmentation retains significant value in the LLM era, particularly when prioritizing real robustness over in-distribution leaderboard performance.

Future work: Expand counterfactual pairs targeting remaining failure modes (emotional manipulation, tu quoque, post hoc); extend to multi-class detection; investigate per-fallacy-type oversampling and advanced generation methods; conduct larger-scale human studies; explore applicability to other reasoning brittleness problems (NLI negation, temporal/causal reasoning).

References

Cal Blanco, Gavin Dsouza, Hugo Lin, and Chelsey Rush. 2025. Slurg: Investigating the feasibility of generating synthetic online fallacious discourse. *Preprint*, arXiv:2504.12466.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *Preprint*, arXiv:2104.02145.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2021. ATOM: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, volume 12975 of *Lecture Notes in Computer Science*, pages 430–445, Cham. Springer International Publishing.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *Preprint*, arXiv:2002.06305.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2024. Missci: Reconstructing fallacies in misrepresented science. *Preprint*, arXiv:2406.03181.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation. *Preprint*, arXiv:2503.23363.

Kyohoon Jin, Juhwan Choi, Jungmin Yun, Junho Lee, Soojin Jang, and Youngbin Kim. 2025. Coba: Counterbias text augmentation for mitigating various spurious correlations via semantic triples. *Preprint*, arXiv:2508.21083.

Zhijing Jin, Yuen Liu, Rada Mihalcea, and Bernhard Schölkopf. 2022. LOGIC: A benchmark for logical reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 790–809, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *Preprint*, arXiv:1909.12434.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Mariia Kutepova and Khalid Al Khatib. 2025. Hybrid intelligence for logical fallacy detection. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 197–208, Suzhou, China. Association for Computational Linguistics.

Abhinav Lalwani, Tasha Kim, Lovish Chopra, Christopher Hahn, Zhijing Jin, and Mrinmaya Sachan. 2025.

Autoformalizing natural language to first-order logic: A case study in logical fallacy detection. *Preprint*, arXiv:2405.02318.

Zhicheng Lin. 2025. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3).

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Preprint*, arXiv:2301.06627.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *Preprint*, arXiv:1907.07355.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2023. How susceptible are llms to logical fallacies? *Preprint*, arXiv:2308.09853.

Mykhailo Poliakov and Nadiya Shvai. 2025. Missynth: Improving missci logical fallacies classification with synthetic data. *Preprint*, arXiv:2510.26345.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Kasidit Sermsri and Teerapong Panboonyuen. 2025. Debiasing large language models in thai political stance detection via counterfactual calibration. *Preprint*, arXiv:2509.21946.

Mehmet Sofi, Matteo Fortier, and Oana Cocarascu. 2022. A robustness evaluation framework for argument mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 171–180, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023. Case-based reasoning with language models for classification of logical fallacies. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5188–5196. International Joint Conferences on Artificial Intelligence Organization. Main Track.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

Appendices

A Training Hyperparameters

All three model variants (**cocopure**, **coco_1x**, **coco_pair**) were trained with the following hyperparameters:

- Learning rate: 2e-5
- Batch size: 32
- Epochs: 5
- Optimizer: AdamW with linear warmup and decay
- Early stopping based on validation loss
- Loss function: Binary cross-entropy

B Qualitative Examples

Representative hard-negative examples from the manual evaluation set

> "Multiple studies have shown that regular exercise improves cardiovascular health, therefore I should start going to the gym more often."
>
> → Baseline: Fallacy (over-flags "studies" + "therefore")
>
> → coco_pair: Valid

Fallacious example (still correctly detected)

> "I got better right after I started taking these vitamins, so they must have cured me."
>
> → coco_pair: Fallacy (post hoc)

Remaining failure example (false negative—appeal to pity)

> "Think of the poor children who will suffer if we don't approve this charity program!"
>
> → coco_pair: Wrongly predicted Valid

C Prompt Templates Used for LLM Evaluation

To ensure full reproducibility and transparency, the exact prompt templates used to evaluate the open-weight LLMs (Llama-3.1-8B and Llama-3.3-70B) on the 300-sample real-world manual evaluation set are provided. All prompts were applied in a zero-shot or few-shot manner without chain-of-thought reasoning, self-consistency, or additional post-processing beyond simple output parsing.

The test argument was inserted directly into the `{text}` placeholder in each template. Model outputs were automatically mapped to binary labels as follows:

- "yes", "fallacy", "FALLACY", "contains fallacy", etc. → fallacious
- "no", "valid", "VALID", "no fallacy", etc. → valid

Minor string normalization was applied to handle minor variations in capitalization and phrasing.

C.1 Zero-shot Engineered (Optimized) Prompt

This carefully designed prompt includes explicit task instructions and output constraints to elicit the strongest possible zero-shot performance.

```
Classify the following argument as either
"VALID" (logically sound reasoning) or
"FALLACY" (contains flawed reasoning or
a logical error).

Argument: {text}

Answer with only one word: VALID or FALLACY
```

C.2 Simple Human-like Prompt

This minimal, natural-language prompt simulates how a non-expert human might ask the question, without task definitions, examples, or output formatting instructions.

Main variant (used for main results):

```
Does this argument have a logical fallacy?
{text}
Answer yes or no:
```

Alternative variant (tested in preliminary experiments, produced comparable degradation):

```
Is this argument valid or does it contain
a logical fallacy?
{text}
Answer: valid or fallacy
```

C.3 Few-shot Prompt (4 Examples)

This prompt provides four human-written in-context examples (two valid, two fallacious) before the test argument, following standard few-shot in-context learning practice.

```
I need help checking whether arguments
contain logical fallacies. Here are some
examples:

Example 1:
"You can't trust his opinion on climate
change because he's not a scientist."
Answer: yes (attacks the person, not
the argument - ad hominem)

Example 2:
"All observed swans in Europe were
white, therefore all swans are white."
Answer: yes (hasty generalization -
does not account for unobserved cases)

Example 3:
"All mammals are warm-blooded. Whales
are mammals. Therefore, whales are
warm-blooded."
Answer: no (valid deductive reasoning)

Example 4:
"The study of 10,000 patients showed
the drug reduced symptoms by 60%.
Therefore, the drug is effective for
this condition."
Answer: no (reasonable evidence-based
inference)

Now evaluate this argument:
"{text}"
Does it have a logical fallacy? Just
answer yes or no:
```

These templates were implemented in the evaluation script. The engineered zero-shot prompt produced the high-performance "optimized" regime

results (F1 90–94%, low FPR). The simple human-like prompts induced the dramatic performance collapse reported in Table 4 (F1 63–72%, FPR 54–74%). The four-shot version achieved partial recovery (F1 83–90%, FPR 21–34%), consistent with known sensitivity of in-context learning to example quality and prompt phrasing.

All prompts are released alongside the evaluation code to facilitate exact replication of the LLM baseline results.