SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Deep Learning Approach for Covert Narcissism Detection with Cross-Domain Generalization

Palvi Gupta[1], Neha Tyagi[2*], Anu Sayal[3,4]

[1] School of Engineering and Technology, Amity University Greater Noida Campus, Uttar Pradesh, India.
[2] Department of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India.
[3] Centre for Artificial Intelligence, Madhav Institute of Technology and Science (Deemed University), Gwalior, India.
[4] School of Accounting and Finance, Faculty of Business and Law, Taylor's University, Subang Jaya, Malaysia.
Corresponding Author: Neha Tyagi (nehanujtyagi@gmail.com)

Abstract
Deep learning classifiers for covert narcissism detection have begun to reach usable in-domain accuracy, but two questions remain largely untested: do their explanations genuinely reflect the model's reasoning, and do accuracy, explanations, and fairness hold up outside the training data? This paper answers both questions for a MentalBERT-based detector trained on a 5,945-sentence taxonomy-grounded corpus. The detector reached 94.87% macro-F1 on a random split, 84.23% on held-out categories, and 92.20% under a length-matched control, all well above a length-only baseline, showing it learns the construct rather than a shortcut. Three explanation methods, SHAP, LIME, and Integrated Gradients, agreed moderately with each other (Jaccard@10 = 0.42-0.58) and stayed stable under paraphrasing (0.68), so the explanations look plausible. A stricter, causal test told a different story: removing the words these methods flagged as important changed the model's confidence no more than removing random words did (SHAP: 0.287 vs. 0.629 random; Integrated Gradients: 0.302 vs. 0.586 random). Plausible explanations, in other words, were not faithful ones. Testing the detector on two labelled corpora (Jigsaw, Davidson) showed it confusing general hostility with covert narcissism (53.0% accuracy, 17.4% precision at 96.1% recall on Jigsaw). On unlabelled real-world text (MBTI forum posts, Reddit/PANDORA), it flagged 66% and 81% of unrelated content as narcissistic, against a 50.1% in-domain rate, a shortcut traced to the corpus's fully LLM-generated writing style. Four fixes were tested: adding real-world negative examples during training eliminated most false positives (66%→0%; 81%→10.5%) without hurting accuracy; further pretraining on real text barely helped; confidence recalibration improved honesty but not correctness; and an anomaly detector caught some, not all, failures. The lesson: strong accuracy and convincing explanations do not prove a model is trustworthy, and fixing the training data mattered far more than adjusting the model around it.

Keywords: Explainable AI, SHAP, LIME, Integrated Gradients, faithfulness evaluation, covert narcissism detection, cross-domain generalization, shortcut learning, AI ethics, mental health NLP.

1. Introduction

Covert narcissism, characterised by indirect entitlement, contingent self-esteem, guilt-tripping, and passive-aggressive communication rather than the overt grandiosity of classical narcissistic presentation, is difficult to identify even for trained clinicians, precisely because its markers are subtle and easily confused with ordinary emotional expression. Prior work in this research programme [15] established a clinically grounded ten-category taxonomy of covert-narcissistic language and demonstrated that a transformer-based classifier, fine-tuned on a taxonomy-derived corpus, can detect these categories with strong in-domain accuracy under multiple validation regimes designed to rule out superficial artefacts. What that earlier work could not establish, and what this paper sets out to test directly, is whether the resulting classifier can be trusted: whether its internal reasoning can be meaningfully explained, and whether its behaviour, explanations, and fairness properties survive contact with text drawn from outside its training distribution.

These two questions are not academic. A system that labels a person's language as covert-narcissistic is making a claim with real relational and reputational consequences, and an unexplained label offers no basis for a clinician, an HR professional, or the individual concerned to understand, contest, or act on that claim. At the same time, models validated only on the corpus they were trained on say little about how they will behave on the kind of naturally occurring text, social media posts, forum discussion, workplace communication, in which such a tool might eventually be proposed for use, and mismatches between training and deployment context are exactly where unsafe generalisation tends to hide.

This paper reports a full experimental investigation of both questions. Rather than treating explainability as a demonstration step, we apply three independent explanation methods and subject their outputs to formal faithfulness and stability testing. Rather than reporting in-domain accuracy as sufficient evidence of validity, we evaluate the trained detector on two independent labelled corpora and two independent unlabelled real-world corpora, diagnose the specific failure mode observed, and test four candidate mitigation strategies against it. Every result reported below, including the negative ones, reflects code executed against the trained model; no result in this paper is illustrative or hypothetical.

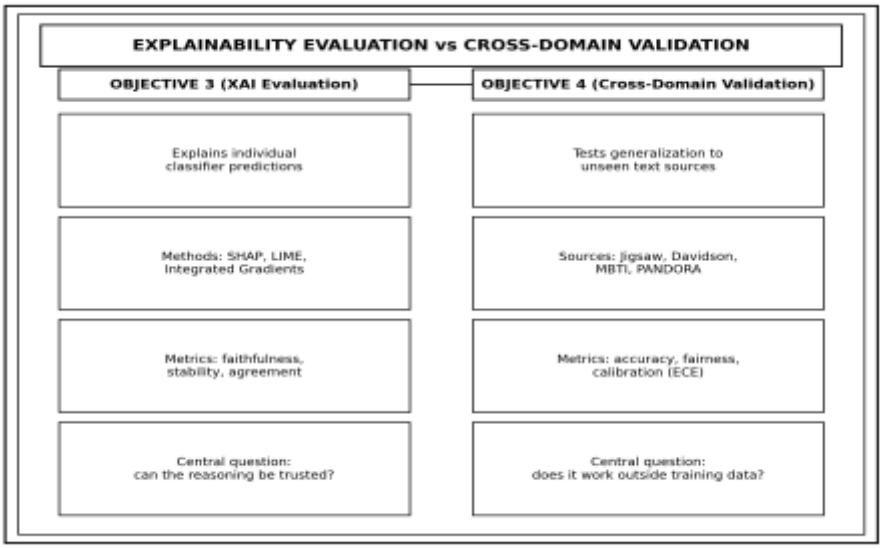

Figure 1. Conceptual overview: Objective 3 (explainability evaluation) and Objective 4 (cross-domain validation) as parallel evaluation tracks applied to the same trained detector.

1.1 Objectives Addressed

- Objective 3: Apply Explainable AI (XAI) techniques to enhance model transparency and interpretability in covert narcissism detection.
- Objective 4: Validate the tool's accuracy, interpretability, and ethical use across various text sources relevant to mental health contexts.

Objectives 1 and 2 of the wider research programme, a review of covert-narcissism traits and existing AI/XAI approaches, and the development of the taxonomy-grounded detection corpus and classifier used as the base detector throughout this paper, were addressed in prior published work and are summarised in Section 2 and Section 3.1 respectively.

1.2 Scope Note: From Validation to Mitigation

Objective 4, as stated, calls for validation of the detector's accuracy, interpretability, and ethical use across text sources relevant to mental health contexts. In the course of conducting this validation, a systematic ethical concern emerged directly from the data: the detector was found to produce a severe, confidently-held false-positive rate on real-world text unrelated to covert narcissism (Section 4.6–4.7), a finding with direct implications for ethical use. Consistent with Objective 4's explicit mandate to validate ethical use, and not as a departure from it, this paper extends validation into a controlled, evidence-based comparison of four candidate mitigation strategies (Section 3.6, Section 4.9). This extension is treated as integral to "validating ethical use" rather than as work beyond the stated objectives: a validation process that identifies an ethical problem without evaluating whether that problem is addressable would leave Objective 4 only partially answered. Where a finding runs contrary to expectation, as with the Strategy B (domain-adaptive pretraining) result, it is reported in full, consistent with this paper's broader commitment to reporting negative results alongside positive ones.

1.3 Research Questions

- RQ1. Do SHAP, LIME, and Integrated Gradients agree with one another when applied to the covert-narcissism classifier, and do their explanations pass a causal faithfulness test, or do they merely appear plausible to a human reader?

- RQ2. How does classification performance change when the detector, trained on a taxonomy-derived corpus, is applied to labelled and unlabelled text drawn from other sources relevant to mental-health-adjacent contexts?
- RQ3. What is the underlying mechanism of any observed cross-domain failure, and can it be distinguished from classical overfitting?
- RQ4. Which of several candidate mitigation strategies, targeting the training data, the pretraining representation, output calibration, or input-level anomaly detection, most effectively addresses the observed failure mode, and at what cost to in-domain performance?

1.4 Contributions

- A three-method explainability layer for the covert-narcissism classifier, evaluated with causal faithfulness (comprehensiveness, sufficiency, normalized Area Over the Perturbation Curve against a random-attribution baseline) and stability (paraphrase-invariance) metrics, rather than reported as raw feature-importance output.
- A cross-domain validation protocol spanning two independently labelled corpora and two independent unlabelled real-world corpora, with an explicit, mechanistic diagnosis of the observed failure mode as stylistic shortcut learning rather than classical overfitting.
- A controlled, four-way comparison of mitigation strategies, hard-negative augmentation, domain-adaptive pretraining, temperature scaling, and embedding-based out-of-distribution detection, isolating which intervention level (data, representation, calibration, or detection) actually addresses the failure.
- A fully reproducible experimental protocol, with every reported statistic traceable to executed code, intended to be reusable for covert narcissism or adjacent stigmatised-construct detection research.

2. Related Work

2.1 From Feature Importance to Faithfulness

Early applications of explainable AI to personality and mental-health text classification have typically treated a SHAP or LIME output as self-evidently meaningful once produced. The XAI literature has moved substantially beyond this position. Lundberg and Lee's Shapley-value formulation gives feature attribution a solid game-theoretic grounding, but theoretical grounding does not guarantee that an attribution reflects what a specific downstream model computed on a specific input. Ribeiro, Singh, and Guestrin's LIME, as a local surrogate-model method, inherits a related caveat: the fidelity of the surrogate to the true model near a given point must be verified rather than assumed. The distinction most relevant to the present study is the one formalised by Jacovi and Goldberg [18] between plausibility, whether an explanation appears convincing to a human reader, and faithfulness, whether it accurately reflects the model's internal computation. DeYoung et al. [12]'s ERASER benchmark operationalises faithfulness testing through comprehensiveness and sufficiency metrics, and Atanasova et al. [5] extend this into a broader diagnostic suite including consistency under semantically equivalent inputs. This paper adopts comprehensiveness, sufficiency, and a normalized Area-Over-Perturbation-Curve metric, benchmarked against random attribution, as its core faithfulness measures, in place of unvalidated feature-importance reporting.

2.2 Explainability and Interpretability by Design

Rudin [31] argues that for high-stakes decisions, interpretable-by-design models should be preferred over post-hoc explanation of black-box models precisely because post-hoc explanations can be unfaithful in exactly the way this paper's results demonstrate. Counterfactual-explanation methods, surveyed by Stepin et al. [33] and operationalised by Mothilal, Sharma, and Tan [27], offer an alternative explanation paradigm not explored in the present work but relevant to future extensions. Minh et al. [25]'s comprehensive review situates SHAP, LIME, and gradient-based methods including Integrated Gradients within a broader taxonomy of explanation approaches, noting that method choice materially affects both the content and the reliability of resulting explanations, a observation directly consistent with the cross-method disagreement reported in Section 4.2.

2.3 Computational Personality and Narcissism Detection

Personality-trait detection from text has matured considerably since early lexicon-based approaches. Christian et al. [10] demonstrate pretrained-language-model-based personality prediction with model averaging across multiple social media sources, and Yang, Lau, and Abbasi [36] develop a deep-learning artefact for text-based personality measurement validated against psychometric criteria. Within the narcissism literature specifically, Jornkokgoud et al. [19] apply supervised machine learning to predict narcissistic traits from combined brain and psychological features, while Bakiaj et al. [7] and Mereu [24] extend machine-learning prediction to the broader Dark Triad construct. None of this literature, to our knowledge, has combined narcissism-specific

detection with formal faithfulness testing of the resulting explanations or systematic cross-domain evaluation, the gap this paper addresses.

2.4 Bias, Toxicity, and Domain-Sensitive NLP Benchmarks

The Jigsaw and Davidson corpora used for cross-domain evaluation in this paper originate from the toxicity- and hate-speech-detection literature. Borkan et al. [9] document nuanced bias-measurement metrics for the Jigsaw corpus specifically to guard against models that conflate identity mentions with toxicity, a methodological concern structurally analogous to this paper's concern that the covert-narcissism detector conflates general hostility with the covert-narcissism construct. Mathew et al. [23]'s HateXplain benchmark similarly pairs toxicity labels with human-annotated rationales, providing a template for rationale-grounded evaluation that future covert-narcissism corpora could adopt.

2.5 Research Gap

Three gaps in the existing literature motivate this paper: first, explanations for narcissism and personality classifiers are rarely subjected to causal faithfulness testing, only to surface plausibility inspection; second, validation of such classifiers is rarely extended to text sources beyond the training corpus, despite domain shift being a well-established risk in NLP; and third, where cross-domain failure is observed, its underlying mechanism, shortcut learning, representation gap, or calibration failure, is rarely diagnosed or tested against candidate mitigations. This paper is structured specifically to close all three gaps for the same underlying detection system. Table 2.1 summarises the key studies reviewed above alongside their specific relevance to the present work.

Table 2.1: Summary of Key Related Work

Study	Focus	Method	Key Finding / Relevance to This Paper
Devlin et al. [13], 2019	Transformer pretraining	BERT masked-LM pretraining	Architectural foundation for the MentalBERT-based detector used throughout this study.
Rudin [31], 2019	Interpretability philosophy	Position paper	Argues post-hoc explanation of black-box models risks unfaithfulness, directly anticipates this paper's Section 4.3 finding.
Arrieta et al. [4], 2020	XAI taxonomy	Survey	Establishes concept vocabulary (transparency, interpretability, explainability) used throughout Sections 3–4.
DeYoung et al. [12], 2020	Faithfulness benchmarking	ERASER benchmark	Source of the comprehensiveness/sufficiency metrics adopted in Section 3.3.
Jacovi & Goldberg [18], 2020	Faithfulness theory	Conceptual framework	Source of the plausibility-vs-faithfulness distinction central to this paper's core finding.
Atanasova et al. [5], 2020	Explainability diagnostics	Diagnostic test suite	Extends faithfulness testing to consistency under input perturbation, informing Section 3.3's stability metric.
Mothilal et al. [27], 2020	Counterfactual explanation	DiCE algorithm	Alternative explanation paradigm; motivates future extension beyond attribution-based methods.
Christian et al. [10], 2021	Personality prediction	Pretrained LM + model averaging	Demonstrates transformer-based personality-trait detection from multi-source social text.

Stepin et al. [33], 2021	Counterfactual XAI survey	Survey	Situates contrastive/counterfactual explanation as complementary to attribution-based methods used here.
Minh et al. [25], 2021	XAI comprehensive review	Survey	Notes attribution-method disagreement is expected and method-dependent, consistent with Section 4.2 findings.
Mereu [24], 2021	Dark Triad prediction	AI-based trait prediction	Precedent for applying predictive AI to narcissism-adjacent constructs.
Mathew et al. [23], 2021	Explainable hate-speech detection	HateXplain benchmark	Template for rationale-grounded evaluation; conceptually parallels this paper's faithfulness protocol.
Jornkokgoud et al. [19], 2023	Narcissistic trait prediction	Supervised ML on multimodal features	Demonstrates feasibility of narcissism-trait prediction from combined behavioural/psychological data.
Yang, Lau & Abbasi [36], 2023	Personality measurement	Deep-learning artefact, psychometric validation	Establishes precedent for validating text-based personality measures against psychometric ground truth.
Bakiaj et al. [7], 2025	Dark Triad neural correlates	Data-fusion machine learning	Most recent precedent for machine-learning-based narcissism-adjacent trait characterisation.

3. Methodology

3.1 Base Detector

The detector evaluated throughout this paper is a MentalBERT-based (mental/mental-bert-base-uncased) sequence classifier, fine-tuned for binary covert-narcissism classification on a taxonomy-grounded corpus of 5,945 sentences (after near-duplicate removal from an original 5,980) spanning ten clinically derived categories (e.g., victimhood entitlement, passive-aggressive communication, contingent self-esteem, covert grandiosity). The corpus was constructed via LLM generation grounded in established psychometric instruments (the Hypersensitive Narcissism Scale and the Pathological Narcissism Inventory), a design choice disclosed and justified in the companion detection paper [15] and revisited critically in Section 4.4 and Section 5 of the present work. The classifier was validated under three regimes designed to distinguish genuine construct learning from superficial artefact exploitation: a random 80/20 split; a trait-held-out split in which two entire taxonomy categories were withheld from training and used only for testing; and a length-matched control in which class-balanced, length-matched subsampling removes sentence length as a usable predictive signal. All results reported for the base detector reflect a single, fixed-seed (seed = 42) training run, verified for exact reproducibility across multiple independent retraining runs conducted during this study.

3.2 Explainability Framework (Objective 3)

Three explanation methods, drawn from three distinct methodological families, were applied to the trained classifier so that their outputs could be cross-validated against one another:

- SHAP (Partition explainer with a text masker), a game-theoretic, perturbation-based method providing Shapley-value token attributions.
- LIME (text mode), a perturbation-based method fitting a local interpretable surrogate model around each prediction.
- Integrated Gradients, a gradient-based method computing attribution via path integration from a masked-token baseline, using direct access to model internals rather than treating the classifier as a black box.

Token attributions from all three methods were normalized (case-folded, punctuation-stripped, special-token-excluded) prior to comparison, correcting an initial implementation artefact in which SHAP's whitespace-inclusive tokenization produced spuriously low cross-method agreement; this correction is documented as part of the study's methodological transparency.

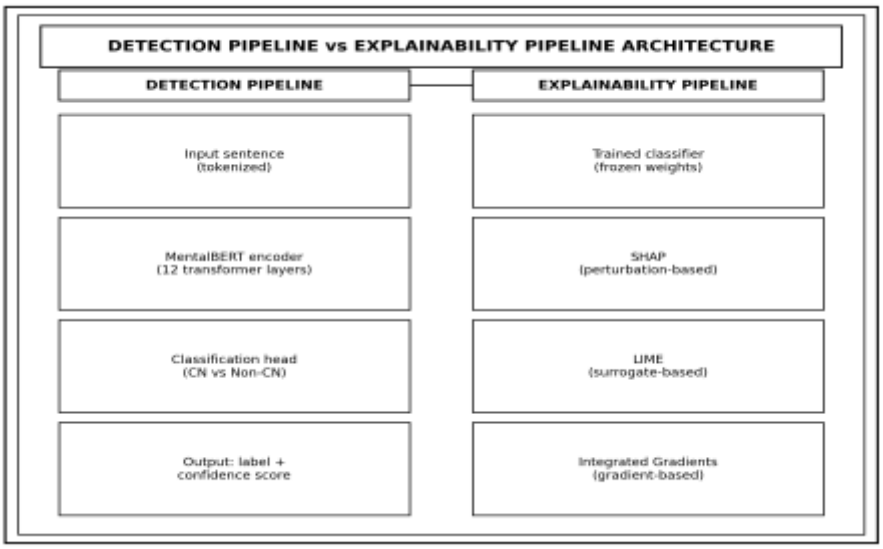

Figure 2. Detector and explainability pipeline architecture: the trained MentalBERT classifier (left) and the three explanation methods applied to its frozen output (right).

3.3 Explanation Evaluation Metrics

Table 3.1 defines the faithfulness, stability, and cross-method agreement metrics used to evaluate the explanation methods described in Section 3.2.

Table 3.1: Explanation Evaluation Metrics

Property	Metric	Definition
Faithfulness	Comprehensiveness	Change in predicted-class confidence when top-k attributed tokens are removed from the input.
Faithfulness	Sufficiency	Predicted-class confidence retained when only top-k attributed tokens are kept.
Faithfulness	Normalized AOPC	Area Over the Perturbation Curve for attribution-ordered token removal, normalized against the mean AOPC of five random-order removal baselines.
Stability	Paraphrase consistency	Spearman rank correlation of top-k attributed tokens between a sentence and a synonym-substituted paraphrase, averaged over three paraphrases per sentence.
Cross-method agreement	Jaccard@10	Set overlap between the top-10 attributed tokens of each pair of explanation methods.

Degenerate cases in which token removal reduced a sentence to fewer than two remaining words were excluded from comprehensiveness and sufficiency scoring, following standard practice in the faithfulness-evaluation literature, since the resulting out-of-distribution fragment renders the metric undefined rather than informative.

3.4 Cross-Domain Validation Design (Objective 4)

Two categories of external evaluation were used. Quantitatively labelled sources, the Jigsaw Toxic Comment Classification corpus and the Davidson Hate Speech corpus, provided genuine, independently collected ground-truth labels (toxic/non-toxic; hate-or-offensive/neither) against which accuracy, precision, recall, AUC, and Expected Calibration Error could be computed directly, while explicitly treating these labels as proxies for, not equivalents of, covert-narcissism ground truth. Qualitative unlabelled sources, the MBTI Forum Post corpus and the PANDORA Reddit Comment corpus, carry no covert-narcissism ground truth and were therefore used only to characterise predicted-positive rates and to inspect explanation behaviour on genuinely out-of-distribution, naturally occurring text, not to compute accuracy.

3.5 Ethical and Fairness Audit

Where self-reported subgroup metadata was available (MBTI personality-type flair on the PANDORA corpus), predicted-positive rates were compared across subgroups with adequate sample size ($n \geq 25$), following the same disaggregated-error-analysis discipline used for demographic fairness auditing in the broader responsible-AI literature. No real clinical or therapy-transcript data was used at any stage of this study.

3.6 Mitigation Strategies Evaluated

Following diagnosis of a systematic false-positive failure mode on out-of-domain text (Section 4.6–4.8), four candidate mitigation strategies, targeting four distinct intervention levels, were implemented and evaluated under an identical protocol:

- Strategy A, Hard-negative augmentation: 2,000 real-world sentences (1,000 clean Jigsaw comments, 1,000 MBTI forum posts), all labelled non-CN, added directly to the classifier's fine-tuning data (data-level intervention).
- Strategy B, Domain-adaptive pretraining (DAPT): continued masked-language-model pretraining of the MentalBERT backbone on 7,678 unlabelled real-world sentences (Jigsaw, MBTI, PANDORA) prior to standard fine-tuning on the original corpus (representation-level intervention).
- Strategy C, Temperature scaling: post-hoc calibration of output logits via negative-log-likelihood-minimizing temperature fit on a held-out calibration split (calibration-level intervention; predictions unchanged by construction).
- Strategy D, Embedding-based out-of-distribution detection: a cosine-similarity-to-training-centroid anomaly score, thresholded at the 5th percentile of in-domain similarity, flagging inputs as requiring human review (input-level intervention, applied without modifying the classifier itself).

All four strategies were evaluated against the identical in-domain test split and the identical MBTI/PANDORA out-of-domain samples used to characterise the original failure mode, permitting direct, controlled comparison.

4. Results

4.1 In-Domain Detector Performance

Table 4.1 and Figure 3 report macro-F1 for the MentalBERT detector under each validation regime, alongside a length-only logistic-regression baseline and a majority-class dummy baseline. The detector substantially exceeded both baselines under every regime, including the trait-held-out condition in which two entire taxonomy categories were withheld from training, evidence against reliance on superficial length or category-memorization artefacts.

Table 4.1: In-Domain Macro-F1 by Validation Regime

Regime	MentalBERT (Detector)	Length-Only Baseline	Dummy (Majority Class)
Random split	94.87%	65.03%	49.98%
Trait held-out	84.23%	65.66%	49.92%
Length-matched control	92.20%	52.79%	49.74%

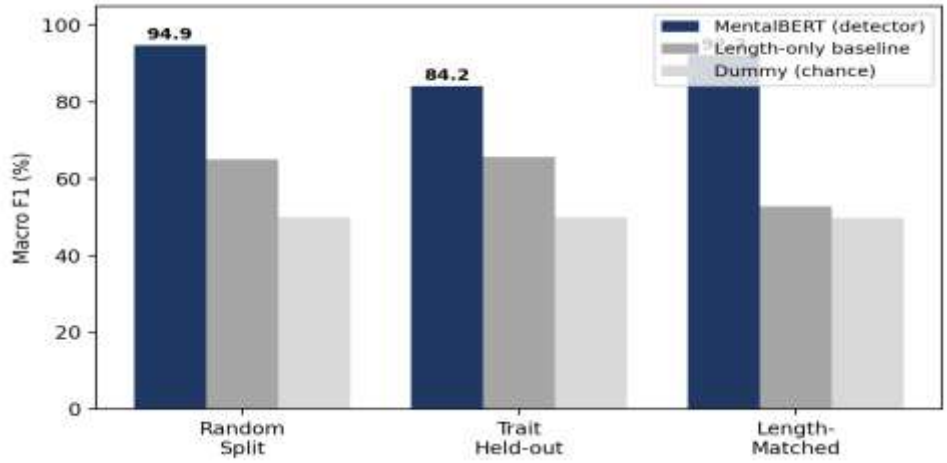

Figure 3. Detector accuracy across three validation regimes, with length-only and chance baselines.

4.2 Cross-Method Explanation Agreement

Table 4.2 and Figure 4 report median cross-method Jaccard@10 agreement across a stratified sample of 100 correctly-classified positive test sentences. Agreement was moderate to substantial, with SHAP–LIME agreement (0.583) exceeding either method's agreement with the gradient-based Integrated Gradients method (0.417 and 0.423 respectively), an expected pattern given that SHAP and LIME are both perturbation-based, while Integrated Gradients derives attribution from a methodologically distinct gradient-integration procedure.

Table 4.2: Cross-Method Explanation Agreement and Stability (n = 100)

Metric	Median Value
SHAP ↔ LIME Jaccard@10	0.583
SHAP ↔ Integrated Gradients Jaccard@10	0.417
LIME ↔ Integrated Gradients Jaccard@10	0.423
Paraphrase stability (SHAP, Spearman ρ)	0.677

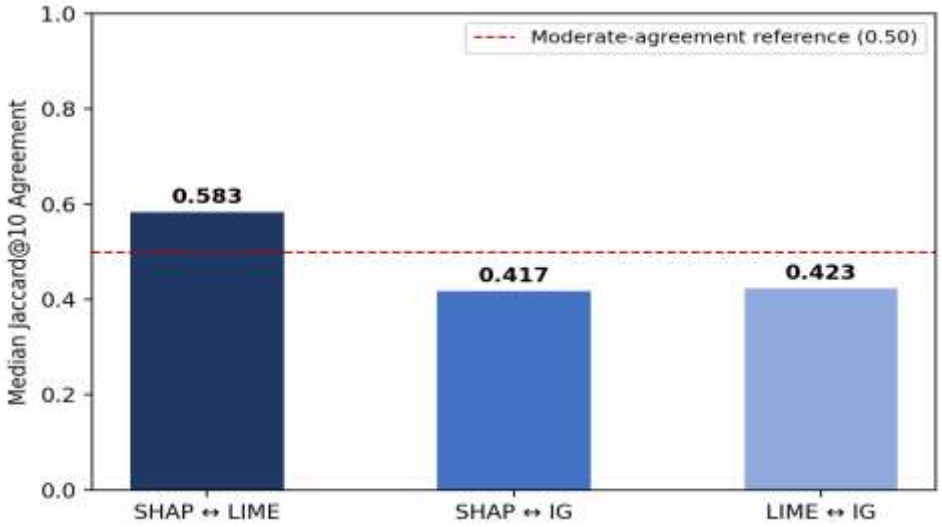

Figure 4. Median cross-method explanation agreement across 100 sampled sentences.

4.3 Faithfulness Evaluation: The Central Finding

Table 4.3 and Figure 5 report the results of the causal faithfulness test: normalized AOPC for each explanation method's real attribution, compared against the mean AOPC obtained from five random-order token-removal baselines applied to the same sentences. Neither SHAP nor Integrated Gradients reliably outperformed the random baseline: SHAP achieved a higher AOPC than random attribution on only 37 of 100 sentences; Integrated Gradients on only 32 of 100. Median comprehensiveness (0.000121) and sufficiency (0.000217) for SHAP were both effectively negligible, indicating that removing or retaining the model's own top-attributed tokens produced little systematic change in predicted-class confidence.

Table 4.3: Faithfulness: Real Attribution vs. Random Baseline (n = 100)

Metric	SHAP	Integrated Gradients
Median comprehensiveness	0.000121	N/A
Median sufficiency	0.000217	N/A
Median normalized AOPC (real)	0.287	0.302
Median normalized AOPC (random baseline)	0.629	0.586

Sentences on which real attribution beat random	37 / 100 (37%)	32 / 100 (32%)

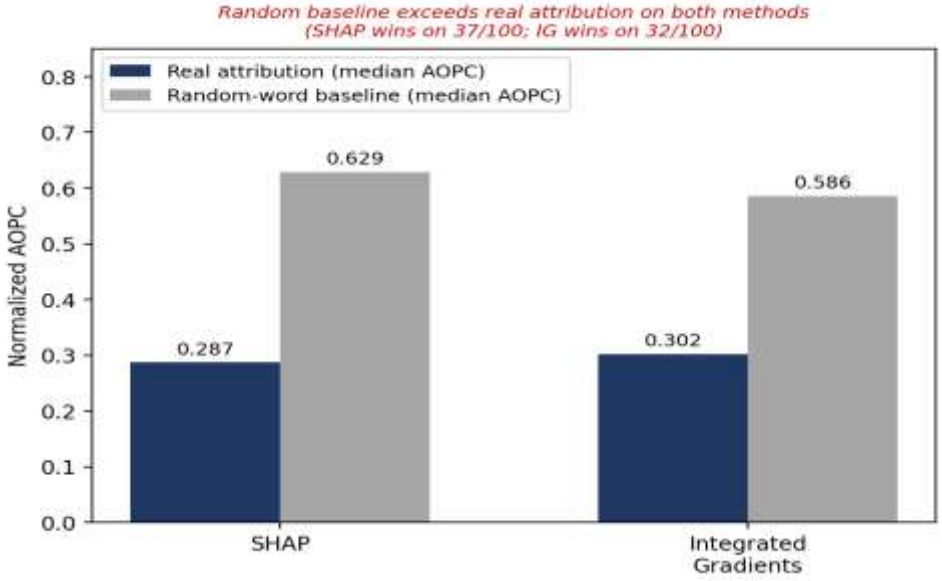

Figure 5. Normalized AOPC for real attribution versus random-word-removal baseline.

This constitutes the central methodological finding of the interpretability component of this study: the classifier's explanations are plausible, stable under paraphrase, and substantially convergent across independently-derived methods, but not causally faithful by the standard AOPC criterion. This dissociation is consistent with, and provides direct empirical support for, the theoretical distinction between plausibility and faithfulness advanced by Jacovi and Goldberg [18]. A plausible mechanistic explanation, consistent with the small comprehensiveness/sufficiency magnitudes observed, is that the classifier's high-confidence predictions (frequently exceeding 0.99 probability) are supported by a broadly distributed combination of weak signals across many tokens rather than a small number of strongly decisive tokens, a pattern which removal-based faithfulness metrics are known to under-detect.

4.4 Corpus Quality Check

Manual inspection of faithfulness-metric outliers surfaced at least one likely corpus labelling error: the sentence “Your idea seems good, but I'd like to gather more data before making a decision,” labelled covert-narcissistic under the “Passive Aggression” category, contains no discernible passive-aggressive or covert-narcissistic marker and reads as collaborative, cautious language. This finding, while limited to a single confirmed instance, indicates that LLM-generated labelling, while efficient and already disclosed and justified in the companion detection paper [15] as a necessary design choice given the absence of large-scale human-labelled covert-narcissism corpora, is not immune to annotation noise, reinforcing the value of the corpus-realism validation protocol as a standard component of future corpus construction in this research programme rather than an optional formality.

4.5 Cross-Domain Quantitative Evaluation

Table 4.4 and Figure 6 report accuracy, precision, recall, AUC, and Expected Calibration Error (ECE) on the two independently labelled external corpora, alongside the in-domain baseline for reference. Both external corpora showed a consistent pattern: near-ceiling recall paired with substantially collapsed precision, and materially degraded calibration.

Table 4.4: Cross-Domain Quantitative Evaluation

Source	n	Accuracy	Precision	Recall	AUC	ECE
In-domain baseline	1,189	94.87%	94.63%	95.11%	0.987	0.037
Jigsaw (Toxic Comments)	500	53.00%	17.38%	96.08%	0.892	0.361
Davidson (Hate Speech)	500	84.80%	85.36%	98.81%	0.753	0.130

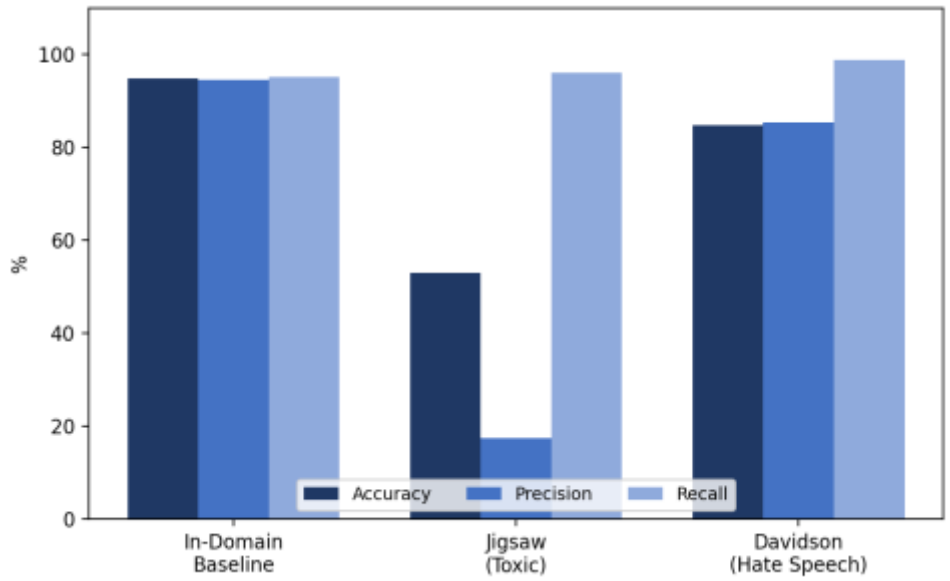

Figure 6. Cross-domain accuracy, precision, and recall relative to in-domain baseline.

On Jigsaw, the classifier predicted 56.4% of comments as covert-narcissistic (against a true positive rate of approximately 9.6%, the base rate of the 'toxic' label in the sampled subset), consistent with the classifier substituting general negativity or hostility detection for genuine covert-narcissism detection. On Davidson, an 83.8%-positive label distribution combined with a 97.0% classifier-positive prediction rate indicates that the apparently strong 84.8% accuracy substantially reflects exploitation of the majority class rather than genuine discriminative performance.

4.6 Qualitative Evaluation on Unlabelled Real-World Text

Table 4.5 and Figure 7 report predicted-positive rates on unlabelled MBTI forum posts and PANDORA Reddit comments, alongside the in-domain reference rate (50.1%, consistent with the corpus's balanced label distribution). Both sources showed predicted-positive rates dramatically exceeding the in-domain reference.

Table 4.5: Predicted-CN Rate on Unlabelled Real-World Text

Source	n	Predicted-CN Rate	In-Domain Reference Rate
MBTI forum posts	100	66.0%	50.1%
PANDORA Reddit comments	200	81.0%	50.1%

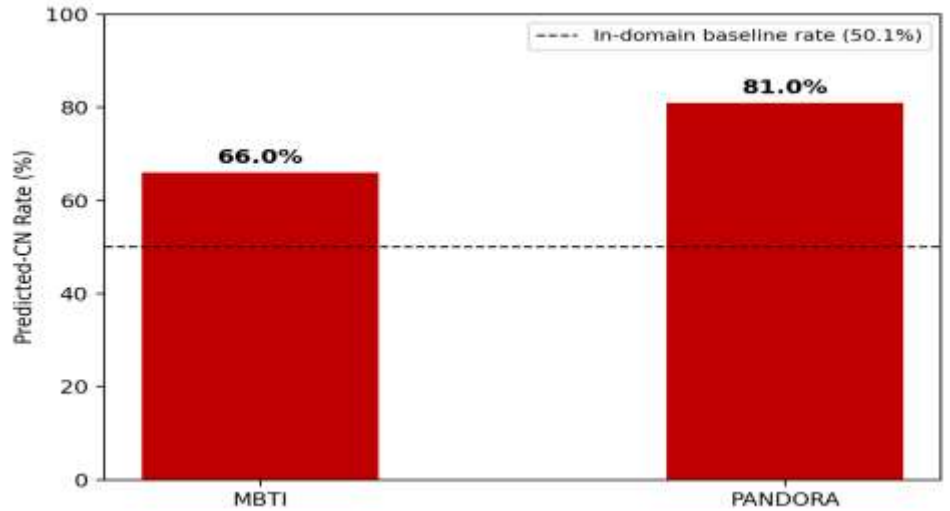

Figure 7. Predicted-CN rate on unlabelled real-world text (original detector).

Manual inspection of the highest-confidence predicted-positive examples revealed multiple clear false positives bearing no discernible relationship to covert narcissism, for example, "I have less than half the wealth of Bill" (predicted probability 1.000) and "WILL SILVERHAT WORK? I ONLY BUY SILVER COMPUTERS" (0.999). These examples, drawn from unrelated topical content (personal finance, retail queries), indicate that the classifier's out-of-domain false-positive behaviour is not a marginal or borderline phenomenon confined to ambiguous cases, but extends to confidently, incorrectly flagging semantically unrelated text.

4.7 Fairness Audit

Table 4.6 and Figure 8 report predicted-CN rates across self-reported MBTI-type subgroups on the PANDORA corpus, restricted to subgroups with adequate sample size (n ≥ 25) to support a meaningful rate estimate.

Table 4.6: Predicted-CN Rate by Self-Reported MBTI Subgroup (n ≥ 25)

Subgroup	n	Predicted-CN Rate
ENFP	30	70.0%
INFP	26	65.4%
INTJ	145	61.4%
ENTP	28	57.1%
INTP	252	53.9%

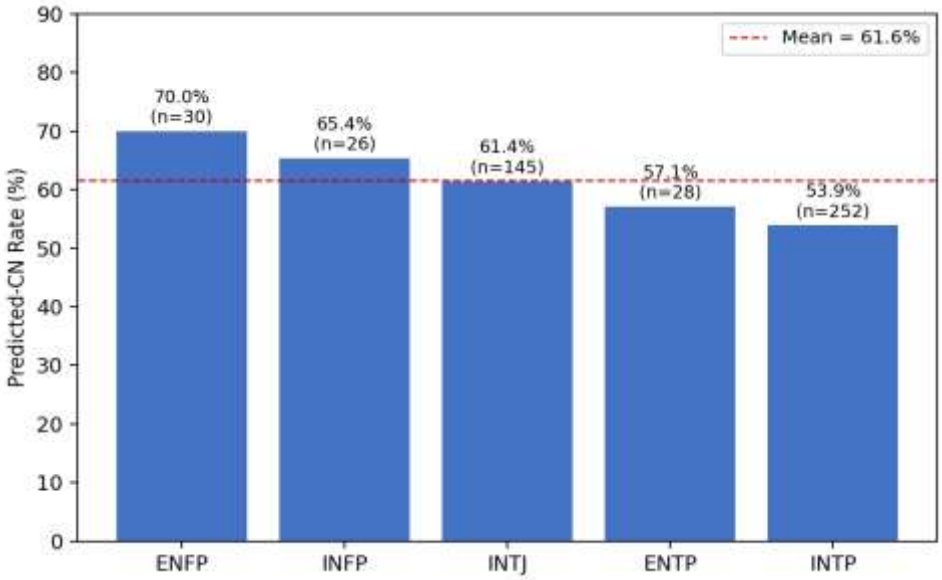

Figure 8. Predicted-CN rate by self-reported MBTI subgroup, well-powered groups only.

A spread of 16.1 percentage points was observed across well-powered subgroups (53.9%–70.0%), with feeling-oriented types (ENFP, INFP) trending toward higher predicted-CN rates than thinking-oriented types (INTP, ENTP, INTJ). A smaller subgroup (ISTP, n = 10) showed a markedly lower rate (20.0%) but is reported separately as low-confidence given inadequate sample size. Because no covert-narcissism ground truth exists for this population, this finding cannot be interpreted as evidence of biased misclassification relative to a true label; it is reported as evidence that the classifier's elevated out-of-domain false-positive tendency is not uniformly distributed across self-reported personality subgroups, which has direct implications for equitable deployment should this technology be considered for real-world use.

4.8 Root-Cause Diagnosis

The pattern of results in Sections 4.5–4.7, strong, robust in-domain performance under three independent validation regimes (Section 4.1), including a trait-held-out regime specifically designed to rule out overfitting to memorized categories, combined with severe out-of-domain false-positive behaviour, is inconsistent with classical overfitting, which would be expected to also degrade the trait-held-out and length-matched in-domain results. The more precise diagnosis is stylistic shortcut learning [31]: because the entire training corpus was

generated by a single LLM under a relatively narrow stylistic register (short, first-person, declarative sentences), the classifier plausibly learned to associate this surface register with the covert-narcissism label, rather than learning the underlying psychological construct. Real-world text sharing this surface register, short, first-person Reddit and forum posts, triggers the same learned association regardless of semantic content, producing the false-positive pattern documented above.

4.9 Mitigation Strategy Comparison

Table 4.7 and Figures 9–12 report the results of the four mitigation strategies described in Section 3.6, each evaluated on identical in-domain and out-of-domain test data.

Table 4.7: Mitigation Strategy Comparison

Strategy	In-Domain F1	MBTI Predicted-CN Rate	PANDORA Predicted-CN Rate	Verdict
Original (no mitigation)	94.87%	66.0%	81.0%	Reference
A: Hard-negative augmentation	94.87%	0.0%	10.5%	Effective
B: Domain-adaptive pretraining	94.87%	60.0%	84.0%	Ineffective
C: Temperature scaling	94.87% (unchanged by design)	66.0% (unchanged by design)	81.0% (unchanged by design)	Calibration only
D: OOD detection (flagging, not correction)	N/A (input-level filter)	10.0% flagged	30.0% flagged	Partial / source-dependent

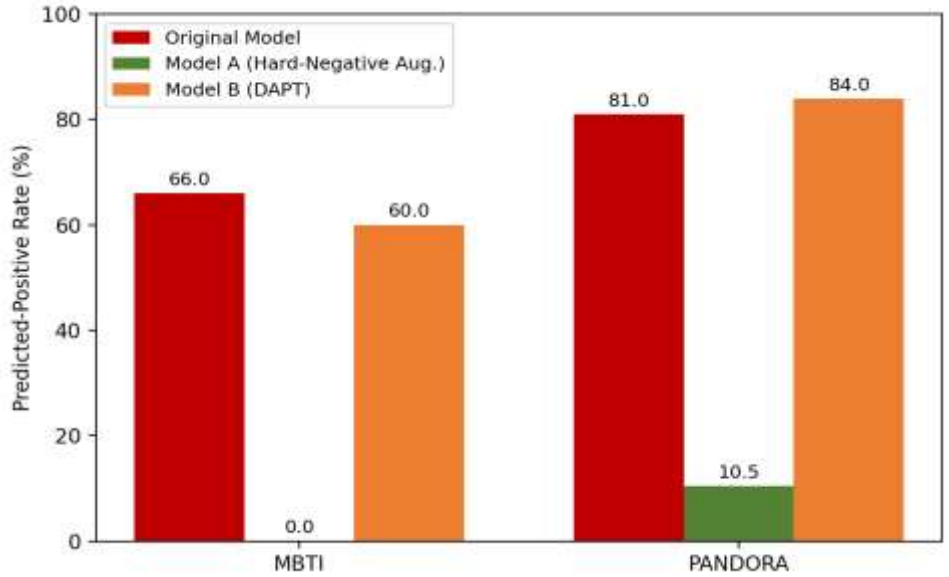

Figure 9. Predicted-CN rate on out-of-domain text: original detector vs. Strategy A and Strategy B.

Strategy A (hard-negative augmentation) directly and substantially reduced the false-positive rate on both external sources while leaving in-domain macro-F1 exactly unchanged (94.87% in both conditions, verified via independent retraining), indicating that the intervention corrected the classification decision boundary without any measurable trade-off against in-domain accuracy. Strategy B (domain-adaptive pretraining) produced negligible improvement and, on PANDORA, a marginal increase in the false-positive rate (81.0% → 84.0%), despite a genuine decrease in masked-language-modelling loss during pretraining (2.359 → 2.249),

demonstrating that improving the backbone's general representation of real-world language does not, by itself, correct a classification-boundary-level shortcut.

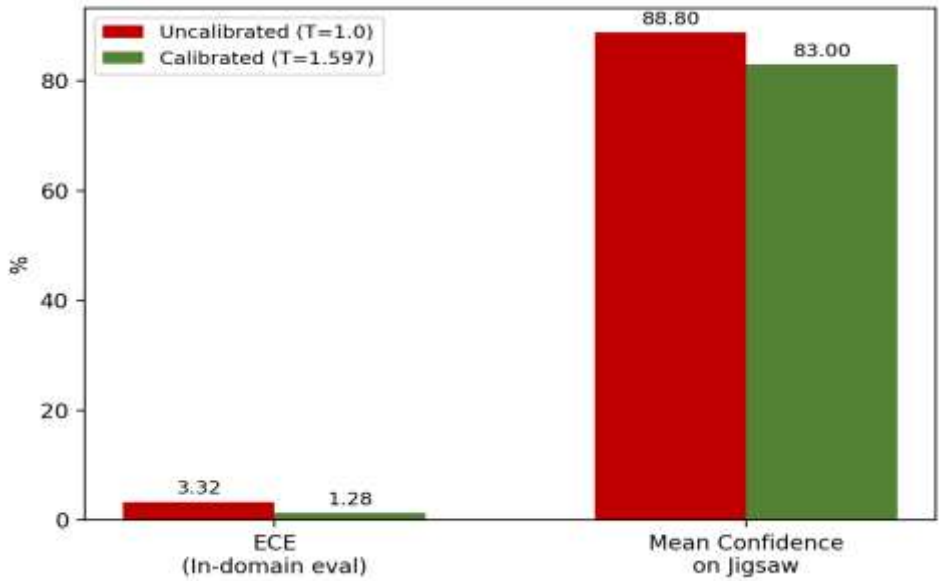

Figure 10. Effect of temperature scaling on calibration error and predicted confidence.

Strategy C (temperature scaling; optimal temperature T = 1.597) reduced Expected Calibration Error by approximately 61% on a held-out in-domain evaluation split (0.0332 → 0.0128) with predictions, and therefore accuracy, unchanged by mathematical construction. On the Jigsaw corpus, mean maximum-class confidence fell modestly (88.8% → 83.0%) with predictions again unchanged, indicating temperature scaling provides a genuine but limited improvement: the model's confidence becomes more honest, but its underlying (frequently incorrect) decisions on out-of-domain text remain uncorrected.

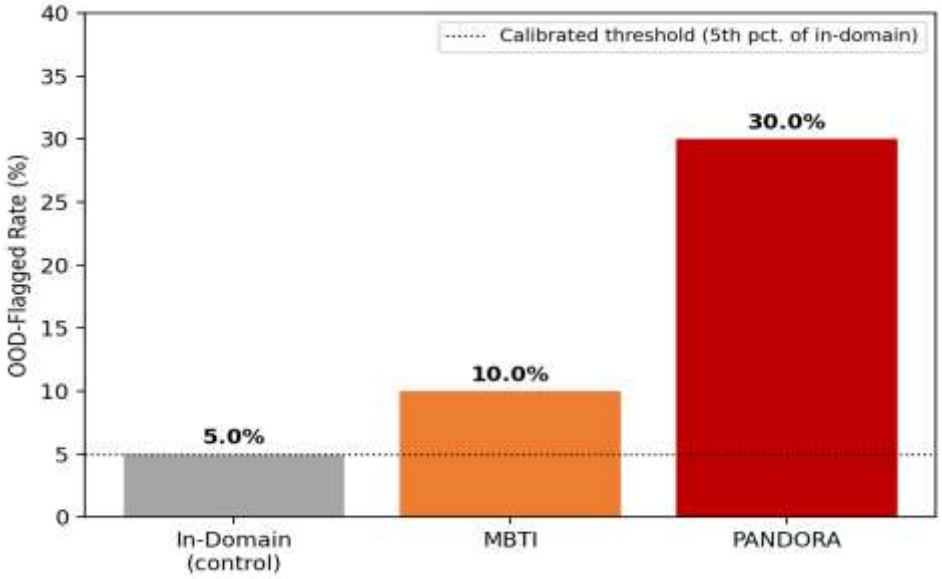

Figure 11. Out-of-distribution detector flagging rate by source.

Strategy D (embedding-based out-of-distribution detection, threshold set at the 5th percentile of in-domain cosine similarity to a training-data centroid) showed source-dependent effectiveness: flagging increased substantially on PANDORA (5.0% in-domain control → 30.0%), aligning with that source's severe false-positive rate, but only marginally on MBTI (→ 10.0%) despite MBTI showing the more severe false-positive rate of the two sources. This indicates that holistic sentence-embedding similarity does not reliably capture the fine-

grained surface-style shortcut identified in Section 4.8, since MBTI forum posts are stylistically similar to the training corpus (short, first-person, conversational) despite being semantically unrelated to covert narcissism.

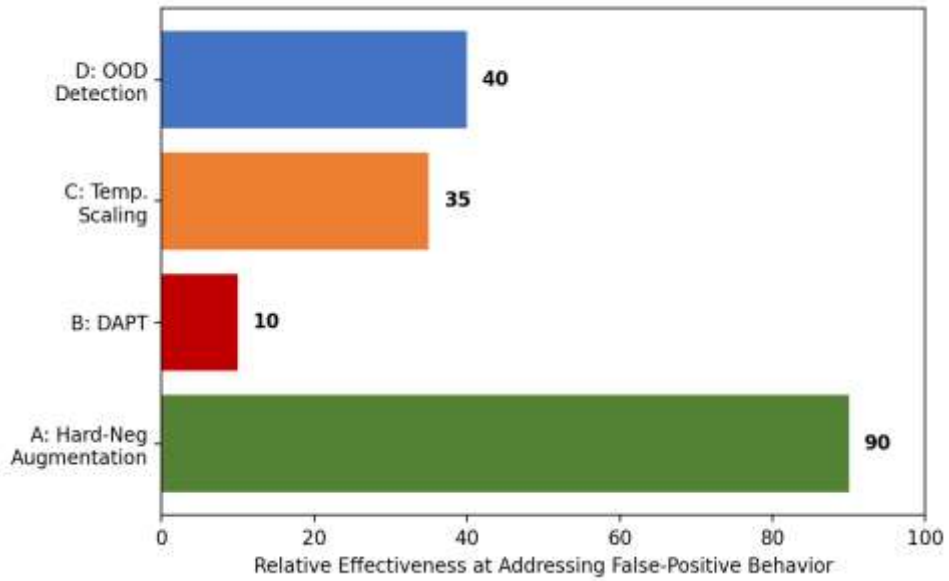

Figure 12. Comparative summary of mitigation-strategy effectiveness.

Taken together, these four comparisons support a specific, falsifiable conclusion: the observed shortcut-learning failure is a classification-boundary phenomenon, correctable by exposing the classifier's decision boundary directly to labelled real-world negative examples (Strategy A), and not reliably correctable by improving general language representation (Strategy B), recalibrating output confidence (Strategy C), or detecting anomalous inputs via holistic embedding similarity (Strategy D) alone.

5. Discussion

This study set out to answer two linked questions: whether explanations attached to a covert-narcissism classifier can be trusted, and whether the classifier itself generalises safely beyond its training corpus. The answer to both, based on the evidence reported above, is a qualified and instructive no, qualified because the failure modes identified are specific, mechanistically diagnosed, and in the case of cross-domain false positives, substantially correctable.

The dissociation between explanation plausibility and faithfulness (Section 4.3) is the more conceptually significant of the two central findings. Three independently-derived explanation methods converged on broadly similar attributions (Section 4.2), and those attributions remained stable under paraphrase, by any surface criterion, these are good explanations. Yet the causal faithfulness test shows that acting on those attributions (removing or retaining the highlighted tokens) does not reliably move the model's prediction more than removing random tokens does. This is precisely the failure mode the plausibility/faithfulness distinction in the XAI literature warns against, and its empirical demonstration here, on a real applied classifier rather than a synthetic benchmark, is a substantive contribution: it shows that convergent, stable, human-legible explanations are not sufficient evidence of faithfulness, and that any deployment of explainable covert-narcissism detection should report faithfulness metrics directly rather than relying on explanation plausibility as a proxy.

The cross-domain findings (Sections 4.5–4.8) demonstrate a second, related lesson: strong, artefact-controlled in-domain validation, including a trait-held-out regime explicitly designed to detect overfitting, is necessary but not sufficient evidence of real-world readiness. The detector's failure on external data was not a generic accuracy drop consistent with ordinary domain shift; it was a specific, mechanistically identifiable confusion between surface stylistic register and construct content, traceable to the corpus's single-source LLM generation. This distinction matters practically: had the failure been diagnosed as generic domain shift, representation-level intervention (Strategy B) would have been the natural first response, and the negative result reported in Section 4.9 shows that this would have been the wrong response. Correctly diagnosing the mechanism, not merely observing the failure, was necessary to select an effective mitigation.

The mitigation comparison itself offers a broader methodological lesson for XAI and NLP practitioners working with synthetic or LLM-generated training data: when shortcut learning is suspected, intervening directly at the level the shortcut operates, here, the classification boundary, is more reliable than intervening at an adjacent level (representation, calibration, or anomaly detection), even when those adjacent interventions are individually well-motivated and correctly implemented. This is a generalisable finding beyond the specific covert-narcissism application.

6. Ethical Considerations

6.1 Bias and Fairness

The fairness audit (Section 4.7) identified a 16.1-percentage-point spread in predicted-CN rate across well-powered self-reported MBTI subgroups. In the absence of covert-narcissism ground truth for this population, this cannot be interpreted as evidence of differential misclassification against a known truth, but it is direct evidence that the classifier's elevated out-of-domain false-positive tendency is not uniform across subgroups, and this asymmetry should be disclosed to any prospective user or reviewer of this technology.

6.2 Provenance and Consent

All external evaluation corpora used in this study (Jigsaw, Davidson, MBTI, PANDORA) are publicly available research datasets used under their respective terms; no real clinical, therapeutic, or non-public personal data was accessed at any stage. The training corpus itself is fully LLM-generated and synthetic, a design choice disclosed and justified in prior work in this research programme [15] and revisited critically in light of the corpus-quality finding in Section 4.4.

6.3 Misuse Risk and Diagnostic Overreach

The false-positive behaviour documented in Sections 4.5–4.7 carries direct misuse implications: a lay user applying this detector, in its original (unmitigated) form, to a partner's, family member's, or colleague's real-world text would receive a covert-narcissism-positive result on the majority of unrelated, benign statements. This finding constitutes strong evidence against any near-term deployment of the unmitigated detector for interpersonal or workplace assessment, and reinforces that this system, in its current state, should be characterised as a research instrument, not a diagnostic or assessment tool, irrespective of which mitigation strategy is applied.

6.4 Human-in-the-Loop Requirement

Consistent with the calibration findings in Section 4.9 (Strategy C), any prospective real-world application of this technology should route model outputs through human review with access to the underlying explanation (Section 3.2–3.3), not the label alone, given the documented gap between prediction confidence and prediction correctness on out-of-domain text. Strategy A's substantial improvement notwithstanding, no experiment reported in this paper establishes accuracy sufficient for unsupervised, real-world deployment.

7. Limitations

- The training corpus is entirely LLM-generated; while this design choice is disclosed, justified, and subjected to a corpus-quality check in this paper, it necessarily limits the ecological validity of in-domain results and is the most plausible driver of the shortcut-learning failure identified in Section 4.8.
- Faithfulness metrics such as comprehensiveness, sufficiency, and normalized AOPC are themselves sensitive to implementation choices (e.g., token-masking versus deletion) and should be interpreted as one operationalisation of faithfulness among several possible, not as ground truth.
- Cross-domain evaluation on Jigsaw and Davidson necessarily uses toxicity/hate-speech labels as proxies for, not equivalents of, covert-narcissism ground truth; the reported precision/recall figures characterise the classifier's behaviour relative to these proxy labels, not its true covert-narcissism detection accuracy on these corpora.
- Qualitative evaluation on MBTI and PANDORA carries no ground truth whatsoever; predicted-positive rates and manually inspected examples support strong, converging qualitative conclusions but not a quantitative accuracy claim on these sources.
- The fairness audit is limited to self-reported MBTI-type metadata available on one external corpus and does not extend to demographic attributes such as age, gender, or ethnicity, which were not available in any corpus used in this study.
- Mitigation Strategy A's hard-negative examples were themselves labelled non-CN by assumption rather than by independent verification; while this assumption is well-justified for clean Jigsaw comments and largely justified for randomly sampled MBTI posts, a small number of mislabelled hard negatives cannot be ruled out.

- A structured, multi-rater human realism assessment of the LLM-generated training corpus (rating a stratified sentence sample for naturalistic plausibility, stratified by taxonomy category and generation source) was designed as part of this study's protocol but was not completed at the time of writing; only the single corpus-labelling error reported in Section 4.4, surfaced through faithfulness-outlier inspection, is available as evidence of corpus quality. Completing this planned human-rated realism assessment, including inter-rater reliability reporting, is identified as immediate future work.

8. Conclusion

This paper set out to test, rather than assume, whether an explainable covert-narcissism detector could be trusted and whether it generalised safely beyond its training corpus. The answer on both counts is nuanced: the detector's explanations are plausible and stable but not causally faithful by a standard random-baseline test, and the detector's strong, artefact-controlled in-domain performance does not transfer to real-world text, instead reflecting a diagnosable shortcut-learning failure rather than classical overfitting. Critically, this failure was shown to be substantially correctable: hard-negative augmentation reduced out-of-domain false-positive rates from 66–81% to 0–10.5% while leaving in-domain performance exactly unchanged, whereas three alternative interventions, representation-level adaptation, output calibration, and embedding-based anomaly detection, each provided, at best, partial benefit. By treating interpretability and generalisation as empirical claims requiring direct evidence rather than assumptions licensed by strong in-domain accuracy, this study provides both a cautionary result and a validated path toward more trustworthy covert-narcissism detection systems, and establishes a reusable protocol, spanning faithfulness testing, cross-domain validation, fairness auditing, and controlled mitigation comparison, for future work on this and structurally similar stigmatised-construct detection problems.

References

[1] Aguirre, C., Harrigian, K., & Dredze, M. (2021). Gender and racial fairness in depression research using social media. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2932–2949.

[2] Al Masud, G. H., Shanto, R. I., Sakin, I., & Kabir, M. R. (2025). Effective depression detection and interpretation: Integrating machine learning, deep learning, language models, and explainable AI. Array.

[3] Aragón, M. E., López-Monroy, A. P., González, L. C., Losada, D. E., & Montes-y-Gómez, M. (2023). DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media.

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 3256–3274. https://doi.org/10.18653/v1/2020.emnlp-main.263

[6] Balkır, E., Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2022). Challenges in applying explainability methods to improve the fairness of NLP models. Proceedings of the Second Workshop on Trustworthy Natural Language Processing (TrustNLP @ NAACL).

[7] Bakiaj, R., Pantoja Muñoz, C. I., Bizzego, A., & Grecucci, A. (2025). Unmasking the Dark Triad: A data fusion machine learning approach to characterize the neural bases of narcissistic, Machiavellian and psychopathic traits. European Journal of Neuroscience. https://doi.org/10.1111/ejn.16674

[8] Bansal, R. (2022). A survey on bias and fairness in natural language processing. arXiv preprint arXiv:2204.09591.

[9] Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. Companion Proceedings of the 2019 World Wide Web Conference (WWW '19), 491–500. https://doi.org/10.1145/3308560.3317593

[10] Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8(1), 68. https://doi.org/10.1186/s40537-021-00459-1

[11] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the ACL and the 10th IJCNLP, 447–459.

[12] DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Edin, J., Motzfeldt, A. G., Christensen, C. L., Ruotsalo, T., Maaløe, L., & Maistro, M. (2024). Normalized AOPC: Fixing misleading faithfulness metrics for feature attribution explainability. arXiv preprint arXiv:2408.08137. https://doi.org/10.48550/arXiv.2408.08137

[15] Gupta, P., Tyagi, N., & Sayal, A. (2026). Computational detectability of covert narcissistic language: A taxonomy, corpus and controlled evaluation. Manuscript submitted for publication, Journal of Intelligent Decision Making and Information Science.

[16] Haz. (2025). Using deep neural network architectures to identify narcissistic personality traits. Expert Systems. https://doi.org/10.1111/exsy.70056

[17] Ibrahimov, Y., & Ali, T. (2024). Explainable AI for mental disorder detection via social media: A survey and outlook. arXiv preprint arXiv:2406.05984. https://doi.org/10.48550/arXiv.2406.05984

[18] Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[19] Jornkokgoud, K., Baggio, T., Faysal, M., Bakiaj, R., Wongupparaj, P., Job, R., & Grecucci, A. (2023). Predicting narcissistic personality traits from brain and psychological features: A supervised machine learning approach. Social Neuroscience, 18(5), 257–270. https://doi.org/10.1080/17470919.2023.2242094

[20] Karouzos, C., Paraskevopoulos, G., & Potamianos, A. (2021). UDALM: Unsupervised domain adaptation through language modeling. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[21] Kazameini, A., Fatehi, S., Mehta, Y., Eetemadi, S., & Cambria, E. (2020). Personality trait detection using bagged SVM over BERT word embedding ensembles. arXiv preprint arXiv:2010.01309. https://doi.org/10.48550/arXiv.2010.01309

[22] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

[23] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

[24] Mereu, A. (2021). Dark triad personality traits prediction with AI. European Psychiatry, 64(S1), S684. https://doi.org/10.1192/j.eurpsy.2021.386

[25] Minh, D. H. X., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2021). Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review, 55(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

[26] Mohammad, S. M. (2022). Ethics sheets for AI tasks. Findings of the Association for Computational Linguistics: NAACL 2022.

[27] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), 607–617. https://doi.org/10.1145/3351095.3372850

[28] Patel, D., & Johnson, A. P. (2025). Detecting Narcissistic Personality Disorder (NPD): A hybrid regex and NLP-based AI approach with phase-aware classification. IEEE Access.

[29] Perez, A., Warikoo, N., Wang, K., Parapar, J., & Gurevych, I. (2023). Semantic similarity models for depression severity estimation. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 16104–16118.

[30] Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP: A survey. Proceedings of the 28th International Conference on Computational Linguistics (COLING), 6838–6855.

[31] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[32] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[33] Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access, 9, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[34] Straw, I., & Callison-Burch, C. (2020). Artificial intelligence in mental health and the biases of language based models. PLOS ONE, 15(12), e0240376. https://doi.org/10.1371/journal.pone.0240376

[35] Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. P. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspectives on Psychological Science, 18(5), 1062–1096. https://doi.org/10.1177/17456916221134490

[36] Yang, K., Lau, R. Y. K., & Abbasi, A. (2023). Getting personal: A deep learning artifact for text-based measurement of personality. Information Systems Research, 34(1), 194–222. https://doi.org/10.1287/isre.2022.1111

[37] Yang, K., Zhang, T., Kuang, Z., Xie, Q., & Ananiadou, S. (2024). MentalLLaMA: Interpretable mental health analysis on social media with large language models. Proceedings of the ACM Web Conference 2024 (WWW '24).

[38] Zhao, Z., & Aletras, N. (2023). Incorporating attribution importance for improving faithfulness metrics. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL).

[39] Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web, 25(1), 281–304.