Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Generative-AI Rewriting and the Robustness of Health Misinformation Detection: A Cross-Dataset Evaluation of Failure Mechanisms and Mitigation

MKULULI SIKOSANA

mkululi.sikosana@stu.mmu.ac.uk

Manchester Metropolitan University

Research Article

Keywords: health misinformation, generative AI, misinformation laundering, COVID-19 FNIR, Monkeypox, detector robustness, register control, computational persuasion features, adversarial robustness

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11507284/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background: Generative AI can alter the surface form, fluency, rhetorical organisation, and evidential register of a health claim without necessarily changing its underlying proposition. High original-text accuracy therefore does not establish model robustness to claim-preserving rewriting, introducing a critical vulnerability into automated public health infrastructure and personalised digital health tools.

Objective: To quantify how ChatGPT-based neutral, ELM central-route, and ELM peripheral-route transformations alter health-misinformation detector performance, computational persuasion features, and detector failure mechanisms, and to evaluate whether transformation-aware training mitigates detector failures and improves the reliability of automated health-information assessment.

Methods: COVID-19 FNIR and Monkeypox corpora were evaluated under a paired C0–C3 design using four classical and five CNN/recurrent detectors. A balanced 200-root intervention cohort was generated with one C1–C3 descendant per condition. For mitigation, each dataset's intervention roots were divided into 50 augmentation roots and 50 untouched mitigation-test roots; C0-only augmentation was compared with C0+C1–C3 transformation-aware augmentation.

Results: Across the initial intervention panel, macro-F1 fell from 0.987 to 0.428–0.479 on FNIR and from 0.922 to 0.731–0.781 on Monkeypox. Failure modes differed sharply, revealing distinct information-quality risks relevant to digital health safeguards. FNIR failure was dominated by false-positive inflation, indicating a risk of incorrectly flagging reliable health information, whereas mpox showed substantial increases in misinformation false negatives, indicating a risk of failing to identify rewritten misleading claims. Transformation-aware training increased mean C1–C3 macro-F1 on unseen Monkeypox roots by 0.148 across nine models (9/9 improved; Wilcoxon p=0.0039) while preserving panel-level C0 performance (+0.004). FNIR mitigation was smaller and heterogeneous (+0.035; 7/9 models improved; p=0.164), and cross-dataset transfer was inconsistent.

Conclusions: Generative-AI rewriting produced substantial, dataset-dependent deterioration in health-misinformation detector performance, even under transformation instructions intended to preserve the underlying claims. Transformation-aware augmentation improved in-domain robustness, particularly for mpox, but did not consistently generalise across datasets. These findings support the inclusion of transformation-based stress testing and class-specific error analysis when evaluating potential automated health-information safeguards. Independent semantic-fidelity validation and application-specific evaluation remain necessary before clinical deployment implications can be established.

1. Introduction

Health-misinformation detection has reached a point at which very high in-domain classification scores are common, but benchmark discrimination and transformation robustness are not the same property. A classifier may reproduce a stable decision boundary on familiar data while remaining sensitive to changes in register, fluency, syntax, rhetorical organisation or emotional tone. Generative AI makes this

distinction operationally important because the surface form of a claim can now be altered rapidly and repeatedly while the proposition remains nominally unchanged.

This study therefore separates two roles for artificial intelligence. Discriminative models classify health content; generative models alter the linguistic realisation of fixed claims. The scientific intervention is the latter. COVID-19 FNIR and Monkeypox are first used to establish original-text competence (C0). Their held-out claim families are then transformed into neutral, central-route and peripheral-route variants (C1-C3). Frozen detectors are applied to all conditions so that the effect of rewriting can be estimated within the same underlying claim family rather than confounded with topic or label changes.

The rationale is strengthened by recent research on misinformation laundering. [1] report that detectors can perform worse on LLM-paraphrased fake news and show that high semantic similarity can coexist with sentiment shift. [2] demonstrate an even more important design issue: when only one class is machine-rewritten, the detector can exploit register; when both real and fake classes are passed through the same laundering process, performance can fall sharply on some benchmarks. The present design therefore transforms both veracity classes symmetrically and treats register control as part of internal validity rather than as an optional robustness check.

This issue is especially relevant to COVID-19 FNIR. The dataset combines false content routed through fact-checking sources with true content from verified publishers [4]. Such construction is valuable for supervised learning but can create source- and register-linked cues. Near-ceiling C0 performance should therefore be interpreted as benchmark predictability, not universal veracity understanding. [3] similarly argue that register variation is a central methodological concern in linguistic fake-news analysis. The transformation framework turns that concern into an experimental test by rewriting both classes under matched conditions.

The Monkeypox corpus provides a second disease context characterised by distinct message lengths, social media conventions, and linguistic profiles. [5] showed that pandemic corpora differ substantially in readability and persuasion-related language. In the framework of personalised medicine, analysing these distinct outbreak environments is critical. As public health communication increasingly relies on digital platforms to deliver tailored health guidance, individuals are exposed to highly variable, unverified online narratives. Automated text-classification guardrails must therefore be capable of processing diverse linguistic structures to ensure that personalised digital health interventions remain grounded in trustworthy, accurate information. By utilising a two-dataset design, this study evaluates whether generative laundering effects present a systemic threat to digital health infrastructure or remain specific to an isolated outbreak corpus.

Relevance to personalised digital healthcare

Personalised digital healthcare increasingly depends on the delivery of health information that is appropriate to an individual's circumstances, information needs and disease-related risks, consistent with international efforts to develop people-centred, digitally enabled healthcare systems [6]. Such

information may be accessed through digital health platforms, patient-facing conversational systems, and online health information services, with emerging evidence demonstrating the potential of conversational AI to support health communication and respond to patient enquiries [7, 8]. When automated misinformation detection is incorporated into these environments, its reliability becomes a key determinant of the quality of information available for individualised prevention and healthcare decision-making, particularly given the documented risks associated with health-related misinformation and the need for responsible AI governance [9, 10]. Generative-AI rewriting introduces an additional challenge because clinically relevant claims may be reformulated into different linguistic and persuasive registers while retaining their underlying meaning. Consequently, misinformation classifiers that perform well on conventional benchmarks may behave inconsistently when processing reformulated health information, as demonstrated by the transformation-induced performance deterioration observed in the present study. The present study investigates this underlying computational vulnerability, providing evidence relevant to the development and evaluation of reliable AI-mediated health-information services, rather than directly evaluating personalised clinical interventions.

The present study is intentionally fully computational. RQ3 and RQ4 analyse measurable text changes and their relationship to detector vulnerability, keeping the intervention, outcomes and mechanism variables at the message/model level. The central premise is therefore that health-misinformation detection should be evaluated not only for discrimination on original text, but also for invariance to alternative linguistic realisations of the same underlying claim. Generative-AI rewriting provides a controlled means of testing that invariance. If detector decisions change substantially after proposition-preserving reformulation, then high benchmark accuracy reflects competence within a particular linguistic regime rather than robust recognition of veracity-related information. The complete workflow linking dataset preparation, C0 qualification, C1-C3 transformation, fidelity control, robustness analysis, computational mechanism analysis and mitigation is summarised in Fig. 1.

2. Related Work

2.1 Health misinformation as a high-stakes computational detection problem

Health misinformation is not simply a generic text-classification problem. False or misleading medical claims can affect treatment choices, preventive behaviour and trust in health institutions, which gives detection errors a different risk profile from many other content-moderation tasks. Health communication research has therefore treated misinformation as both a public health problem and an information quality problem [9, 11]. Systematic reviews show that misinformation is recurrent across vaccines, infectious diseases, drugs, non-communicable diseases, and medical treatments, with social media providing a major route of exposure [12]. COVID-19 intensified this problem and stimulated a large body of research on automatic detection, but the resulting literature is heterogeneous in data sources, labelling policies, features, and evaluation protocols [13, 14].

That heterogeneity matters for model interpretation. [14] found that health-misinformation studies span traditional machine learning, deep learning and transformer-based approaches, but also identified limited multilingual and cross-domain evaluation. [13] similarly concluded that COVID-19 misinformation detection had developed rapidly around benchmark classification while leaving open challenges in reproducibility, feature selection and generalisation. These reviews motivate the present emphasis on robustness rather than another isolated accuracy comparison. A detector intended for health information must remain reliable when language, source conventions, and presentation change, because these properties are unstable in real-world information environments.

2.2 Detection architectures and benchmark competence

The broader fake-news literature has long combined lexical, stylistic, user and contextual signals [15]. In text-only settings, sparse linear classifiers remain strong controls because they expose how much of a benchmark can be separated through lexical regularities alone. Neural models add different inductive biases: 1D-CNNs emphasise local phrase patterns [16], LSTMs model sequential dependencies [17], BiLSTMs use bidirectional context [18], and CNN-recurrent hybrids combine local and sequential composition.

Pandemic misinformation benchmarks demonstrate how high in-domain performance can become. Patwa et al. [19] released a manually annotated COVID-19 fake-news dataset and showed that conventional supervised models could already achieve strong discrimination. [20] later demonstrated that transformer-based COVID-19 models, particularly domain-adapted encoders combined with recurrent layers, can reach very high F1 performance. In the COVID-19 FNIR, [21] reported strong conventional, neural and hybrid results, while subsequent ELM-informed modelling further increased performance [22]. These results establish benchmark competence, but they do not establish that the learned decision function is invariant to semantically faithful changes in wording.

This distinction is central to the present design. The detector ladder is deliberately broad because robustness may depend on representation. Sparse linear classifiers, convolutional neural networks and recurrent or hybrid neural architectures may exhibit different vulnerabilities under the same linguistic transformation. The scientific endpoint is therefore not which architecture wins C0, but whether competence survives controlled rewriting and whether any mitigation transfers across model families and datasets.

2.3 Theory-informed linguistic features in health-misinformation detection

Health misinformation is also characterised by how claims are framed. The Elaboration Likelihood Model distinguishes central-route processing, associated with scrutiny of arguments and evidence, from peripheral-route processing, which relies more heavily on cues such as affect, authority, or salience [23]. The Theory of Planned Behaviour links attitudes, subjective norms and perceived behavioural control to

behavioural intention [24]. In computational text studies, these theories cannot be treated as direct measures of cognition, but they can motivate transparent families of linguistic features.

[25] explicitly used ELM to organise central and peripheral features for health-misinformation detection, integrating linguistic, topical, sentiment, and behavioural information. [22] similarly operationalised readability, sentiment and heuristic cues within an ELM-informed CNN-LSTM framework. Cross-pandemic analysis then showed that COVID-19, Constraint and Monkeypox corpora differ materially in readability and persuasion-related language [5]. More recent multi-branch modelling combines semantic, rhetorical, stance, ELM and TPB-inspired text features in a unified representation [26]. The present study uses these constructs differently: not as unobserved psychological states, but as measurable textual dimensions whose displacement under Generative-AI rewriting can be quantified and related to detector error.

2.4 Generative AI changes the misinformation threat model

Large language models alter the threat model by generating or reformulating fluent content at scale. [27] showed that contemporary LLMs can produce convincing articles aligned with disinformation narratives, while [28] describe LLMs as a dual-use technology that can both amplify the risks of misinformation and support detection or mitigation. A subsequent scoping review reached a similar conclusion, finding evidence for both misinformation generation and defensive uses of Generative AI, while emphasising inconsistent safeguards and the need for standardised evaluation [29].

For detector evaluation, the most relevant issue is not only the generation of wholly new false content but the transformation of existing claims. [30] showed that fake-news detection must account for human-written, machine-generated and machine-paraphrased content rather than treating machine authorship as synonymous with falsity. [1] directly examined LLM-paraphrased fake news and found that paraphrasing can reduce detector performance, with explanation analyses implicating changes such as sentiment shift. [2] then demonstrated why register control is essential: when real and fake classes are both transformed into a common LLM register, some benchmark detectors deteriorate sharply. Together, these studies show that machine register can be either a shortcut or a source of distribution shift, depending on how the experiment is constructed.

The present transformation design also connects to adversarial NLP. [31] introduced semantically equivalent adversarial rules to expose model brittleness under meaning-preserving changes and showed that augmentation with such cases can reduce discovered failures. [32] used syntactically controlled paraphrases to create adversarial examples that alter structure while maintaining paraphrase quality. [33] demonstrated that contextual substitution can also create adversarial text for modern classifiers. These studies establish a general principle: prediction stability under semantically constrained perturbation is a distinct property from held-out accuracy.

However, semantic preservation cannot be assumed from a generation prompt. [34] showed that many purported natural-language adversarial examples violate semantic or grammatical constraints, and that

stricter validity criteria can substantially reduce apparent attack success. The TextAttack framework consequently formalises attacks in terms of a goal, transformations, constraints and search procedures, while also supporting adversarial training [35]. Behavioural testing work reaches the same methodological conclusion from a different direction: standard held-out accuracy can miss systematic failures that become visible only when specific linguistic capabilities are tested [36]. These findings justify the fidelity gate and paired C0-C3 design used here. A detector failure is informative only if the transformed item still represents the source claim closely enough for the comparison to remain valid.

2.5 Dataset bias, calibration and cross-domain generalisation

A second major literature question is whether near-ceiling fake-news scores reflect general veracity inference or dataset-specific shortcuts. [37] identified selection bias and source overlap in unreliable-news datasets, showing that models can exploit source-label mappings rather than the intended task. [3] similarly foreground register as a methodological confound in fake-news analysis. More recent work confirms that this problem persists: [38] show that detection performance depends strongly on dataset construction and annotation granularity, while [39] report that structurally homogeneous benchmarks can support high in-dataset accuracy yet transfer poorly across datasets.

Out-of-distribution evaluation therefore needs to be explicit rather than inferred from an ordinary random test split. [40] showed that fake-news models lose performance when domains differ and proposed methods that preserve both domain-specific and cross-domain information. [41] formalised misinformation generalisation across time, event, topic, publisher, political bias and misinformation type, showing that standard classification metrics can obscure failures under distribution shift. A large multi-dataset comparison by [42] reaches a related conclusion across traditional machine learning, deep learning, transformers and LLM-based approaches: in-domain performance and cross-domain generalisation are different evaluation problems.

Calibration is part of the same issue. [43] showed that strong discriminative models can nevertheless be poorly calibrated. Under linguistic distribution shift, ranking performance and a fixed operating threshold can therefore diverge. This is directly relevant to a transformation study because a model may retain substantial AUROC while producing an unacceptable false-positive or false-negative rate at the threshold selected on C0.

Data augmentation has long been used to reduce text classifiers' dependence on narrow surface forms. [44] demonstrated that simple text augmentation can improve classification, especially with limited data, while adversarial paraphrase work showed that controlled syntactic or semantic variants can be used for robustness training [32, 31]. TextAttack further operationalised adversarial training as a systematic defence against perturbation-induced failure [35]. In the LLM era, [30] argue that detector training distributions must explicitly account for machine-generated and machine-paraphrased content. [45] extend this logic to a broader proactive-defence perspective, in which robustness is built in before deployment rather than treated solely as a post hoc detection problem.

Generative AI therefore has a dual role in the present study. It is the source of the controlled perturbation used to expose brittleness and the mechanism for constructing transformation-aware augmentation for mitigation. The important empirical question is whether that defensive exposure creates invariance that survives unseen source roots and, more stringently, a different health dataset.

2.6 Reliability of AI-mediated health information in personalised digital healthcare

The increasing integration of artificial intelligence into digital healthcare has created opportunities for personalised health-information delivery, patient engagement and healthcare decision support. Conversational agents and automated symptom checkers can improve access to health information and facilitate interactions between patients and healthcare services [48, 49]. However, the reliability and acceptability of these technologies remain important considerations. [46] identified concerns regarding privacy, empathy and technological limitations, while Semigran et al. [47] demonstrated substantial variation in the diagnostic and triage accuracy of online symptom checkers. These findings highlight the importance of evaluating AI-supported healthcare technologies beyond their functional capabilities.

The reliability of automated health-information assessment is particularly relevant where digital platforms provide information adapted to individual health needs. Incorrectly identifying reliable information as misinformation may restrict access to trustworthy health guidance, whereas failing to detect misleading claims may expose users to potentially harmful information [9, 11]. Generative AI introduces additional challenges because health claims can be reformulated into different linguistic and persuasive registers without necessarily changing their underlying propositions. Consequently, strong classification performance on conventional benchmarks may not guarantee robustness when health information is linguistically transformed.

[10, 50] emphasises reliability, transparency, safety and appropriate governance in healthcare AI. Consistent with these principles, the present study evaluates the robustness of health-misinformation detectors under controlled generative-AI transformations, examining classification errors, linguistic changes and mitigation strategies. Its contribution concerns the computational reliability of potential health-information safeguards rather than the direct evaluation of personalised clinical interventions or patient outcomes.

The literature establishes strong health-misinformation classifiers, evidence of dataset and register bias, adversarial brittleness in NLP, and growing research on misinformation in the large language model (LLM) era. Studies of AI-mediated healthcare further highlight the importance of information reliability, patient trust and appropriate safeguards in digital health applications. However, the intersection of these research areas remains insufficiently examined, particularly regarding the robustness of automated health-information assessment when generative AI alters the linguistic and persuasive presentation of health claims. Specifically, limited evidence exists from paired, root-disjoint, register-controlled evaluations that transform both veracity classes symmetrically, distinguish neutral from theory-informed

transformations, quantify transformation-induced linguistic changes, examine their associations with detector failure, and evaluate mitigation on unseen source roots and across health datasets.

This study addresses these gaps using COVID-19 FNIR and the Monkeypox dataset, four classical machine-learning models and five task-trained neural architectures. It combines fixed C0 operating thresholds, C1–C3 ChatGPT transformations, computational persuasion-feature analysis, class-specific error evaluation and root-disjoint transformation-aware mitigation. Beyond assessing classification robustness, the study identifies distinct information-quality failure modes, including the incorrect flagging of reliable health information and the failure to detect rewritten misinformation. These findings provide computational evidence relevant to evaluating potential misinformation-detection safeguards in personalised digital healthcare. The contribution is therefore a systematic investigation of health-misinformation detector robustness under generative-AI transformation, rather than a demonstration of clinical effectiveness or the superiority of any individual classification architecture.

3. Research Questions and Hypotheses

RQ1. Does claim-preserving Generative-AI rewriting alter the classification performance and ranking discrimination of health-misinformation detectors relative to unmodified claims?

RQ2. Do neutral, ELM central-route and ELM peripheral-route transformations differ in their effects across detector families and datasets?

RQ3. How do C1-C3 transformations change measurable linguistic, rhetorical and persuasion-related features relative to the original C0 text, and are those shifts stable across veracity classes and datasets?

RQ4. Which transformation-induced feature shifts are associated with detector confidence loss and misclassification, and do those associations differ by detector or dataset?

RQ5. Can transformation-aware training reduce robustness gaps on unseen claim families and across the held-out health dataset while preserving original-text performance?

H1 predicts a nonzero paired change in detector performance between C0 and at least one transformed condition. H2 predicts condition-by-model and condition-by-dataset interactions. H3 predicts distinct linguistic and persuasion-feature profiles for C1-C3, with possible dataset- and veracity-specific heterogeneity. H4 predicts that within-root feature shifts explain a nontrivial proportion of transformation-induced detector confidence loss and error. H5 predicts that transformation-aware training will reduce robustness gaps on unseen roots and across the held-out dataset without materially degrading C0 performance. Figure 1 maps these hypotheses to the corresponding computational stages and outputs.

4. Materials and Methods

4.1 Datasets and task definition

COVID19-FNIR is a COVID-19-specific corpus containing fact-checked false information and verified true material from news publishers [4]. The supplied corpus contains 7,588 records before duplicate controls. After case-folding, whitespace-normalisation, and exact-text deduplication, the locked baseline corpus contains 7,249 unique records: 3,459 misinformation items and 3,790 reliable items. The binary target is misinformation = 1 and reliable information = 0.

The supplied Monkeypox corpus contains 5,787 rows and 5,266 unique texts. The primary task uses the strict 0-versus-1 veracity subset (n = 2,559) because the residual code-9 category is heterogeneous. The strict task, therefore, provides a cleaner misinformation-versus-information comparison for the transformation study. A secondary sensitivity analysis may reintroduce the residual class according to the original binary mapping, but it is not part of the primary endpoint. Table 1 summarises the two analytic corpora, label definitions and their roles in the study.

Table 1
Dataset composition and role in the fully computational transformation framework.

Dataset	Raw n	Primary analysis n	Primary label	Role
COVID-19 FNIR	7,588	7,249 unique	Reliable vs misinformation	Primary C0 baseline; C1-C3 transformation stress test; cross-dataset mitigation
Monkeypox	5,787	5,266 unique; 2,559 strict 0/1	Informational vs misinformation	Second-disease C0 baseline; C1-C3 stress test; held-out-domain mitigation

4.2 Duplicate control and source-root-disjoint evaluation

Exact duplicate removal was performed before splitting. COVID-19 FNIR was normalised and exact-text deduplicated to 7,249 unique records; the Monkeypox strict task used 2,559 unique code-0/code-1 posts. The locked baseline split used stratified 70/15/15 train-validation-test partitions with random state 42. For the exploratory intervention reported here, 100 roots were sampled from each locked test partition, balanced 50:50 by veracity. Each C1-C3 descendant remained attached to its source root, and no transformed test descendant entered detector training. Although exact-text deduplication was performed before partitioning and each transformed descendant remained associated with its original source root, semantic near-duplicate clustering and comprehensive claim-family overlap detection were not implemented across the original dataset partitions. Accordingly, the evaluation establishes source-root separation for generated transformations but does not guarantee complete semantic independence between training, validation and test claims. The possibility of differently worded instances of equivalent claims occurring across partitions remains a limitation, particularly when interpreting near-ceiling original-text classification performance.

The intervention cohort therefore contains 200 independent source roots and 800 texts: 200 C0 originals plus one C1, C2 and C3 descendant for every root. This paired design makes each source item its own control while preserving the locked detector training data. Vocabulary construction and threshold/model selection used training/validation data only; the transformation cohort was not used to fit or retune the C0 detectors.

4.3 Detector ladder and baseline qualification

The detector architectures evaluated in this study are summarised in Table 2. The executed intervention and mitigation experiments comprised nine models: four classical machine-learning classifiers and five task-trained neural/hybrid architectures. These models were evaluated under the original C0 condition and the C1–C3 Generative-AI transformation conditions.

Table 2
Classical machine-learning and neural/hybrid architectures evaluated in the C0–C3 Generative-AI transformation study.

Family	Model	Representation	Role
Classical	Logistic regression	Word + character TF-IDF	Linear probabilistic control
Classical	Linear SVM	Word + character TF-IDF	High-margin lexical control
Classical	Multinomial Naive Bayes	TF-IDF	Sparse generative control
Classical	Random forest	TF-IDF-derived features	Nonlinear classical comparator
Neural	1D-CNN	Task-trained embeddings	Local phrase sensitivity
Neural	LSTM	Task-trained embeddings	Sequential dependency modelling
Neural	BiLSTM	Task-trained embeddings	Bidirectional sequence context
Neural hybrid	CNN-LSTM	Convolution + recurrence	Local + sequential composition
Neural hybrid	CNN-BiLSTM	Convolution + bidirectional recurrence	Local + bidirectional composition

C0 qualification is not the scientific endpoint; it is a safeguard against floor effects. The primary transformation analysis was restricted to models achieving the prespecified C0 qualification threshold of macro-F1 ≥ 0.90 on the relevant dataset. Model inclusion was determined using macro-F1 on the full original-text held-out test partition (Table 5), rather than the smaller 100-root intervention cohort (Tables 7 and 8). Models below that threshold remain in supplementary sensitivity analysis but are not used to make strong laundering claims. For the executed neural/hybrid models, random seeds,

vocabulary construction, maximum input sequence lengths and checkpoint-selection criteria were fixed before scoring the transformed test texts.

4.4 Generative-AI transformation conditions

The four experimental conditions and the permitted/prohibited rewrite operations are specified in Table 3. C0 is the unmodified control, while C1-C3 are generated descendants of the same source root.

Table 3
Generative-AI intervention conditions.

Condition	Purpose	Permitted changes	Prohibited changes
C0 Original	Control condition	None	Any rewriting
C1 Neutral rewrite	Improve fluency/readability without persuasive reframing	Grammar, syntax, sentence flow, neutral register	New facts/evidence; changed certainty; altered causal or clinical claims
C2 ELM central-route rewrite	Increase argument organisation and explicit reasoning using only existing content	Logical ordering, connective structure, clarity of already-present evidence	Fabricated sources/statistics/citations; new clinical claims
C3 ELM peripheral-route rewrite	Alter bounded surface/peripheral cues while preserving proposition	Rhetorical emphasis, certainty/urgency and peripheral salience while preserving proposition	New facts/evidence; altered clinical proposition; fabricated authority or calls to action

Transformations were generated in ChatGPT using GPT-5.6 Sol (OpenAI; accessed 11 October 2026). For each of 200 locked source roots, one C1, one C2 and one C3 rewrite was generated. C1 requested a neutral fluency/readability rewrite without persuasive reframing or new facts. C2 requested clearer reasoning, evidence organisation and explanation using only information already present in the source. C3 requested a bounded peripheral-route presentation that could alter rhetorical emphasis, urgency, social salience and certainty-related wording while preserving the underlying proposition and label. The same transformation logic was applied to both veracity classes. No transformed text was used to retrain the detector models reported in RQ1-RQ4.

4.5 Transformation fidelity and equivalence gate

Generation prompts explicitly prohibited new medical claims, fabricated evidence, changes in causal direction and changes in veracity. A structured post-generation audit checked source-descendant pairing and immutable-field consistency. Retrospective automated fidelity screening was performed on all 600 generated transformations using lexical similarity, numerical consistency and negation consistency checks. Numerical consistency ranged from 74% to 77% across C1–C3 in COVID-19 FNIR and from 81% to 82% in mpox. Negation consistency ranged from 85% to 91% and from 77% to 79%, respectively. These measures flag potential differences between original and rewritten texts but cannot establish

whether such differences represent clinically meaningful changes in the underlying propositions. The results therefore constitute preliminary computational quality checks rather than independent confirmation of semantic equivalence. Independent pretrained natural language inference, semantic-equivalence assessment and external adjudication were not performed in the executed experiment. Consequently, although the transformation prompts prohibited substantive changes to the underlying health claims and structured checks were applied to source-descendant pairs, complete preservation of clinical meaning and veracity cannot be independently confirmed. The reported C1–C3 results are therefore interpreted as exploratory estimates of detector sensitivity to fidelity-constrained generative-AI rewriting rather than definitive effects of independently verified meaning-preserving transformations. Future confirmatory evaluation should incorporate bidirectional entailment, clinical-claim consistency checks and independent adjudication before transformed items are included in the robustness analysis. Raw transformed misinformation is not reproduced in the main text.

4.6 RQ1-RQ2: detector robustness analysis

For detector m and condition c, the Laundering Robustness Gap was calculated as LRG(m, c) = macro-F1(m, C0) - macro-F1(m, c), where positive values indicate deterioration after rewriting. RQ1 reports macro-F1, AUROC, AUPRC, false-positive rate (FPR) and misinformation false-negative rate (FNR). RQ2 compares C1-C3 across the model panel using Friedman tests within the dataset and Wilcoxon signed-rank contrasts with Holm correction. The same locked decision threshold selected before the transformation stress test was retained across C0-C3, so decision-boundary sensitivity is part of the robustness outcome rather than being removed by post-intervention threshold tuning.

RQ1 reports the observed model-by-condition performance metrics and Laundering Robustness Gaps from the locked intervention cohort. RQ2 uses the executed Friedman omnibus tests and Wilcoxon signed-rank pairwise contrasts with Holm correction across the model panel. No bootstrap confidence interval, mixed-effects estimate or GEE interaction result is reported for RQ1-RQ2 beyond these executed analyses.

Health-information reliability interpretation

To examine the relevance of the computational findings to digital health information services, false-positive and false-negative rates were interpreted as distinct information-quality failure modes. False positives represent reliable health-information items incorrectly classified as misinformation, whereas misinformation false negatives represent misleading items incorrectly classified as reliable. These outcomes provide safety-relevant indicators for the evaluation of potential automated health-information filtering components. The study does not assume that classification errors directly produce adverse clinical outcomes, nor does it evaluate the operation of these models within a deployed personalised healthcare system.

4.7 RQ3: computational linguistic and persuasion-feature analysis

Table 4
Implemented computational feature inventory used in RQ3 and RQ4. Every variable listed was extracted from the C0-C3 intervention texts; no unimplemented or hypothetical feature is included.

Feature family	Implemented variables	Operational role
Readability/structure	word_count; avg_sentence_length; type_token_ratio; FKGL	Length, lexical diversity and readability
Punctuation/typography	question_rate; exclamation_rate; uppercase_ratio; punctuation_rate	Surface emphasis and typographic salience
Social / directive language	pronoun_social_rate; imperative_rate; normative_social_rate	Social reference, directives and normative wording
Central-route evidence/reasoning	evidence_rate; argument_rate; attribution_rate; causal_rate	Evidence, attribution and explicit reasoning cues
Affect / peripheral salience	urgency_rate; negative_affect_rate; positive_affect_rate; authority_rate	Affective, urgency and authority-linked cues
Epistemic framing	certainty_rate	Lexical certainty markers

The implemented feature inventory is reported in Table 4. For each C1-C3 descendant, every listed feature was extracted and differenced from its paired C0 source. Rates were normalised by word count where applicable. RQ3 used paired Wilcoxon tests against zero within condition with false-discovery-rate correction. Additional Mann-Whitney tests compared the resulting within-root deltas across datasets and across veracity classes, with FDR correction within each comparison family. These variables are text-derived operationalisations of rhetorical and ELM/TPB-inspired constructs; they are not measurements of human cognition or behaviour.

4.8 RQ4: computational vulnerability-mechanism analysis

RQ4 modelled transformation-induced vulnerability using paired feature deltas. Instance-level calibrated probabilities were retained for logistic regression and multinomial Naïve Bayes, which form the probabilistic mechanism sample reported here. A cluster-robust binomial model predicted new errors (correct at C0, wrong after transformation), and a cluster-robust linear model predicted confidence loss. Predictors were standardised within the pooled intervention dataset, and models were adjusted for transformation condition, detector identity, and dataset. Interaction terms between each standardised feature delta and dataset, and between each feature delta and detector identity, were tested to assess whether the observed associations were invariant across these strata.

The mechanism analysis is associative rather than causal: a significant feature coefficient indicates that an observed within-root linguistic change covaries with error or confidence loss after adjustment, not that the feature alone caused the failure. False-discovery-rate correction was applied to the supplementary detector- and dataset-interaction tests. Because comparable instance-level probability traces were not retained for the full neural panel, no neural-family mechanism coefficient is reported.

4.9 RQ5: transformation-aware mitigation and generalisation

For RQ5, the 100 intervention source roots within each dataset were stratified by veracity and split with random state 2026 into 50 augmentation roots and 50 untouched mitigation-test roots. The original dataset training and validation partitions remained unchanged. To isolate the effect of transformation-aware augmentation from the effect of adding new source roots, the control regime added only the C0 original text for the 50 augmentation roots to the original training set, whereas the transformation-aware regime added C0 + C1+C2 + C3 for those same roots. Decision thresholds and neural checkpoint selection used only the original validation partition; no mitigation-test descendant informed training, vocabulary construction, threshold selection or early stopping. Both regimes were evaluated on all C0-C3 descendants of the 50 untouched mitigation-test roots. The primary endpoints were the change in mean C1-C3 macro-F1, the reduction in mean LRG, and the change in C0 macro-F1. Cross-dataset transfer was evaluated separately by training/augmenting on one health corpus and testing on the other corpus's untouched mitigation roots. The nested mitigation analysis is reported as an exploratory secondary analysis because it was conducted after the initial RQ1-RQ4 intervention run.

The completed RQ5 analyses therefore cover two forms of generalisation: unseen roots within the same dataset and transfer to the other health dataset. Held-out generator-family evaluation was not performed because the intervention used a single generator, ChatGPT GPT-5.6 Sol; no result for generator-family transfer is claimed.

4.10 Reproducibility, governance and responsible release

Reproducibility materials retained for the reported analyses include the locked split manifests, model configurations, ChatGPT transformation prompts, condition labels, aggregate and instance-level analysis outputs, feature definitions, statistical code and random seeds. Responsible release is based on source identifiers or hashes, transformation metadata, and derived features, rather than on a ready-to-use corpus of rewritten false health claims. The generator is recorded as ChatGPT GPT-5.6 Sol, OpenAI, accessed 11 October 2026; interactive generation did not expose API temperature or top-p controls, which limits exact regeneration.

The computational implementation, reproducibility scripts, frozen aggregate results and associated verification materials are available through the accompanying Zenodo software archive [51].

5. Results

5.1 C0 detector qualification

C0 qualification establishes that the evaluated detectors are competent on unmodified text before the intervention. As shown in Table 5, COVID-19 FNIR macro-F1 ranged from 0.991 to 0.999 across the nine models evaluated, with AUROC ranging from 0.996 to 1.000. Monkeypox was more difficult: macro-F1

ranged from 0.910 to 0.960 and AUROC from 0.971 to 0.996. The C0 results therefore provide a strong pre-intervention reference and show that the two corpora differ materially in baseline difficulty. These values are not evidence of transformation robustness.

Table 5
Locked C0 original-text qualification results on the full held-out test partitions.

Model	FNIR Acc	FNIR macro-F1	FNIR AUROC	Mpox Acc	Mpox macro-F1	Mpox AUROC
Logistic regression	0.999	0.999	1.000	0.953	0.952	0.994
Linear SVM	0.999	0.999	1.000	0.961	0.960	0.996
Multinomial NB	0.991	0.991	1.000	0.956	0.954	0.990
Random forest	0.998	0.998	1.000	0.935	0.932	0.982
1D-CNN	0.996	0.996	1.000	0.911	0.910	0.983
LSTM	0.993	0.993	0.996	0.922	0.919	0.976
BiLSTM	0.997	0.997	1.000	0.922	0.919	0.981
CNN-LSTM	0.998	0.998	1.000	0.924	0.922	0.971
CNN-BiLSTM	0.999	0.999	1.000	0.914	0.912	0.976

5.2 ChatGPT Generative-AI intervention cohort

The intervention analysis used 200 locked test roots: 100 COVID-19 FNIR items and 100 Monkeypox items, with 50 misinformation and 50 reliable/informational roots per dataset. ChatGPT generated one descendant under each of C1-C3, producing 600 transformed texts and 800 condition-specific texts including C0. The condition definitions are given in Table 3, and their position in the end-to-end workflow is shown in Fig. 1. Detector thresholds were not re-optimised after transformation, so each C1-C3 score measures how a C0-calibrated detector responds when the same source proposition is re-expressed.

The transformations were evaluated under the prompt constraints and structured source-descendant checks described in Section 4.5. Neither independent pretrained NLI nor independent adjudication was applied. The C1-C3 estimates are therefore reported as fidelity-constrained exploratory results, and this limitation is carried through the interpretation of the intervention effects.

5.3 RQ1: detector robustness under C1-C3 transformation

Generative-AI transformation produced large, dataset-specific changes in detector behaviour (Table 6). On COVID-19 FNIR, mean macro-F1 fell from 0.987 at C0 to 0.445 at C1, 0.428 at C2 and 0.479 at C3. Mean AUROC declined from 0.998 to 0.871–0.888 and mean AUPRC from 0.998 to 0.890–0.895, indicating that ranking discrimination weakened but remained substantially better than the thresholded classification results. The dominant error was false-positive inflation: the mean FPR increased from

0.022 at C0 to 0.871, 0.891, and 0.833 under C1-C3, while the misinformation FNR remained between 0.000 and 0.007. The principal FNIR failure is therefore a decision-boundary/register shift that causes transformed reliable content to be over-flagged, rather than widespread escape of transformed misinformation.

Monkeypox showed a different failure mode (Table 6). Mean macro-F1 fell from 0.922 at C0 to 0.731, 0.781 and 0.753 under C1–C3; AUROC fell from 0.981 to 0.882, 0.937 and 0.891, and AUPRC to 0.862, 0.927 and 0.887. In contrast to FNIR, the principal safety-relevant change was misinformation false negatives: the mean FNR rose from 0.076 at C0 to 0.409 under C1, 0.276 under C2 and 0.344 under C3. The transformation therefore caused many misinformation items to cross the fixed decision boundary and be missed. These findings distinguish two computational information-quality risks: incorrectly flagging reliable health information and failing to identify rewritten misinformation. However, differences in dataset construction, linguistic characteristics and annotation practices prevent these contrasting patterns from being attributed solely to disease context.

Table 6
Aggregate detector response across the nine executed classical/neural models in the 200-root intervention cohort.

Dataset	Condition	Mean macro-F1	Mean AUROC	Mean FPR	Mean FNR	Mean AUPRC
COVID-19 FNIR	C0	0.987	0.998	0.022	0.004	0.998
COVID-19 FNIR	C1	0.445	0.871	0.871	0.007	0.890
COVID-19 FNIR	C2	0.428	0.882	0.891	0.000	0.893
COVID-19 FNIR	C3	0.479	0.888	0.833	0.000	0.895
Monkeypox	C0	0.922	0.981	0.080	0.076	0.981
Monkeypox	C1	0.731	0.882	0.076	0.409	0.862
Monkeypox	C2	0.781	0.937	0.140	0.276	0.927
Monkeypox	C3	0.753	0.891	0.109	0.344	0.887

Table 7
COVID-19 FNIR macro-F1 by intervention condition; LRG = C0 macro-F1 minus transformed-condition macro-F1.

Model	C0	C1	C2	C3	Mean LRG
Logistic regression	1.000	0.454	0.417	0.417	0.571
Linear SVM	1.000	0.540	0.540	0.586	0.445
Multinomial NB	0.980	0.759	0.757	0.825	0.200
Random forest	1.000	0.507	0.472	0.524	0.499
1D-CNN	0.990	0.333	0.333	0.417	0.629
LSTM	0.990	0.329	0.333	0.397	0.637
BiLSTM	0.980	0.355	0.333	0.376	0.625
CNN-LSTM	0.960	0.391	0.333	0.417	0.580
CNN-BiLSTM	0.980	0.333	0.333	0.355	0.639

Table 8
Monkeypox macro-F1 by intervention condition; LRG = C0 macro-F1 minus transformed-condition macro-F1.

Model	C0	C1	C2	C3	Mean LRG
Logistic regression	0.940	0.848	0.793	0.900	0.093
Linear SVM	0.940	0.807	0.782	0.838	0.131
Multinomial NB	0.940	0.847	0.837	0.817	0.106
Random forest	0.930	0.880	0.839	0.890	0.060
1D-CNN	0.950	0.859	0.830	0.830	0.110
LSTM	0.890	0.448	0.525	0.532	0.388
BiLSTM	0.910	0.828	0.721	0.785	0.132
CNN-LSTM	0.900	0.441	0.860	0.448	0.317
CNN-BiLSTM	0.900	0.622	0.838	0.737	0.167

The model-level results in Tables 7 and 8, visualised in Figs. 2 and 3, reinforce this contrast. On FNIR, Multinomial Naive Bayes was the least affected model by mean LRG (0.200), whereas CNN-BiLSTM (0.639), LSTM (0.637), 1D-CNN (0.629) and BiLSTM (0.625) showed the largest average losses. On Monkeypox, Random Forest had the smallest mean LRG (0.060) and LSTM the largest (0.388); CNN-LSTM was strongly condition-dependent, falling to 0.441 under C1 and 0.448 under C3 but retaining 0.860 under C2. Averaged across models, the mean LRG for FNIR was 0.542, 0.559, and 0.507 for C1-C3,

compared with 0.191, 0.142, and 0.169 for Monkeypox. These results show that the same aggregate performance drop can reflect distinct error pathways and architectural sensitivities.

5.4 RQ2: differences among transformation types

Transformation condition had a statistically detectable but non-uniform effect across the model panel (Table 9). Across all 18 model-dataset pairs, the Friedman test was significant (chi-square = 10.42, p = 0.0055), but none of the pooled C1-C3 pairwise comparisons remained significant after Holm correction. This rules out a universal ordering in which one rewrite type is consistently the most damaging.

Dataset-specific contrasts in Table 9 were more informative. FNIR showed a significant omnibus condition effect (Friedman p = 0.0027), and C2 produced lower macro-F1 than C3 after Holm correction (adjusted p = 0.0234). Monkeypox showed no reliable omnibus difference among C1-C3 (p = 0.4895). RQ2 is therefore supported as a condition-by-dataset phenomenon rather than as a global ranking of neutral, central-route and peripheral-route rewriting.

Table 9
RQ2 transformation-condition contrasts across the executed model panel.

Scope	Contrast	Mean difference	Raw p	Holm p
Pooled	C1 vs C2	-0.016	0.3343	0.3343
Pooled	C1 vs C3	-0.028	0.0304	0.0911
Pooled	C2 vs C3	-0.012	0.0626	0.1253
Pooled	Friedman C1/C2/C3		0.0055	
COVID-19 FNIR	C1 vs C2	0.017	0.0938	0.0938
COVID-19 FNIR	C1 vs C3	-0.034	0.0391	0.0781
COVID-19 FNIR	C2 vs C3	-0.051	0.0078	0.0234
COVID-19 FNIR	Friedman C1/C2/C3		0.0027	
Monkeypox	C1 vs C2	-0.049	1.0000	1.0000
Monkeypox	C1 vs C3	-0.022	0.3008	0.9023
Monkeypox	C2 vs C3	0.028	0.8438	1.0000
Monkeypox	Friedman C1/C2/C3		0.4895	

5.5 RQ3: computational linguistic and persuasion-feature shifts

The transformation conditions produced distinguishable computational feature signatures (Table 10; Fig. 4). All three conditions shortened texts on average while increasing average sentence length and FKGL. C2 produced the largest FKGL increase (+ 5.167 grade levels) and evidence-cue increase (+

0.0108), consistent with a more explicit explanatory/evidential presentation. C3 produced the largest certainty (+ 0.0229), urgency (+ 0.0085) and normative/social (+ 0.0080) shifts. C1 shortened text by 7.12 words on average and increased evidence cues by 0.0091, confirming that the neutral rewrite condition was not linguistically inert.

Table 10
Mean within-root feature shifts relative to C0. * FDR-adjusted p < 0.05.

Feature	C1 delta	C2 delta	C3 delta
Word count	-7.1200*	-4.7700*	-8.4150*
Avg sentence length	+ 5.3801*	+ 7.7176*	+ 4.1134*
FKGL	+ 2.7573*	+ 5.1670*	+ 1.7675*
Evidence rate	+ 0.0091*	+ 0.0108*	-0.0001
Certainty rate	+ 0.0053*	-0.0003	+ 0.0229*
Urgency rate	-0.0008	-0.0014	+ 0.0085*
Normative/social rate	+ 0.0048*	+ 0.0059*	+ 0.0080*

The RQ3 stability tests showed that these shifts were not fully invariant across the corpus or with respect to veracity. Under C2, the FKGL increase was larger for Monkeypox (+ 6.294) than for FNIR (+ 4.040; FDR-adjusted p = 0.0040), while the C1 certainty-rate shift was positive for FNIR (+ 0.0129) but slightly negative for Monkeypox (-0.0023; adjusted p = 0.0497). Veracity also moderated several changes: under C3, shortening was larger for reliable/informational roots (-11.07 words) than for misinformation roots (-5.76; adjusted p = 0.0047), and under C1 the evidence-rate increase was larger for reliable/informational roots (+ 0.0175) than for misinformation roots (+ 0.0007; adjusted p = 0.0168). The simple argument-marker lexicon did not increase significantly under C2 after multiplicity correction. Thus, the prompt-defined conditions created measurable but multidimensional feature profiles rather than perfectly isolated theoretical constructs.

5.6 RQ4: computational mechanisms of detector vulnerability

The pooled mechanism models are reported in Table 11. Across transformed items, new-error rates for the two probabilistic detectors were 0.325, 0.340 and 0.305 on FNIR C1-C3, compared with 0.110, 0.145 and 0.085 on Monkeypox. After adjustment for condition, detector and dataset, positive word-count change was associated with lower odds of a new error (OR = 0.412, 95% CI 0.304–0.556, p < 0.001), meaning that stronger compression was associated with greater vulnerability. Increased FKGL was associated with higher odds of a new error (OR = 1.643, 95% CI 1.133–2.382, p = 0.0088). The corresponding confidence-loss coefficients were − 0.110 for word-count change (p < 0.001) and + 0.050 for FKGL (p = 0.0005).

Table 11
RQ4 mechanism models. Predictors are standardised feature deltas; models adjust for condition, detector and dataset.

Predictor	New-error OR	95% CI	p	Confidence-loss beta	p
Delta word_count	0.412	0.304–0.556	< 0.001	-0.110	< 0.001
Delta fkgl	1.643	1.133–2.382	0.0088	+ 0.050	< 0.001
Delta type_token_ratio	0.900	0.705–1.150	0.3998	-0.001	0.9410
Delta evidence_rate	1.026	0.843–1.250	0.7955	+ 0.011	0.2062
Delta certainty_rate	0.785	0.572–1.076	0.1325	-0.017	0.2176
Delta urgency_rate	0.969	0.813–1.155	0.7284	-0.001	0.8609
Delta normative_social_rate	0.932	0.709–1.225	0.6130	+ 0.002	0.8590

The remaining tested predictors in Table 11 were not independently significant in the pooled new-error model after adjustment. Supplementary interaction tests showed that the mechanism was not completely invariant: the word-count association differed by dataset (FDR-adjusted $p < 0.001$) and by detector (adjusted $p = 0.0085$), while type-token-ratio interactions were significant for both dataset and detector (adjusted $p = 0.0062$ and 0.0085, respectively). Dataset interactions were also detected for certainty and normative/social deltas in the new-error model (adjusted $p = 0.0259$ and 0.0363). These results narrow the interpretation of RQ4: structural compression and readability change are the strongest pooled correlates of vulnerability, but their effect sizes vary across corpora and detectors, and no causal psychological mechanism is claimed.

5.7 RQ5: transformation-aware mitigation and generalisation

On COVID-19 FNIR, transformation-aware training produced a small and heterogeneous change in robustness (Table 12; Fig. 5). Mean C1-C3 macro-F1 increased by 0.035 across the nine-model panel (median + 0.041), with seven of nine models improving; the model-panel shift was not statistically significant (Wilcoxon signed-rank $p = 0.1641$). The largest gains were BiLSTM (+ 0.147), logistic regression (+ 0.105), LSTM (+ 0.086) and CNN-LSTM (+ 0.056). Linear SVM deteriorated by 0.152 and Multinomial Naive Bayes by 0.014. C0 macro-F1 remained 1.000 for every FNIR model in the 50-root mitigation subset, so the average transformed-text gain did not involve a C0 performance trade-off.

Monkeypox showed a substantially stronger and more consistent mitigation effect (Table 12; Fig. 5). All nine models improved on transformed text, increasing mean C1-C3 macro-F1 by 0.148 (median + 0.128; Wilcoxon p = 0.0039). Mean LRG fell by 0.143, while panel-level C0 macro-F1 changed by only + 0.004. The largest transformed-text gains were LSTM (+ 0.250), Multinomial Naive Bayes (+ 0.218), CNN-LSTM (+ 0.204) and CNN-BiLSTM (+ 0.168). Model-level C0 changes were nevertheless heterogeneous: BiLSTM increased by 0.101, whereas Linear SVM and CNN-BiLSTM each decreased by 0.040 and LSTM by 0.022. LRG reduction, therefore, needs to be interpreted jointly with the C0 change rather than treated as a stand-alone improvement metric.

A supplementary class-specific error analysis was conducted for the four classical classifiers using the original nested mitigation design (Supplementary Table S1). Across conditions C1–C3, transformation-aware augmentation reduced the average false-negative rate in the mpox cohort from 27.0% to 1.3%, while the corresponding false-positive rate increased from 5.0% to 6.3%. In COVID-19 FNIR, the average false-positive rate decreased from 65.7% to 63.7%, with no change in the average false-negative rate, which remained at zero. Paired source-root bootstrap analyses based on 2,000 resamples of the 50 held-out roots per dataset indicated positive transformed-text macro-F1 gains for all four classical classifiers in mpox, although the findings were mixed in FNIR. These analyses support the value of transformation-aware augmentation for reducing missed misinformation in the mpox evaluation setting, while also demonstrating that improvements may be model- and dataset-dependent. The bootstrap intervals quantify held-out-root sampling uncertainty conditional on the fixed trained models and do not capture training-seed or generator variability.

Table 12

RQ5 leakage-controlled nested mitigation results on 50 held-out source roots per dataset. Positive Δ transformed F1 and positive LRG reduction indicate improved robustness. Neural results are from seed 42; classical models are deterministic under the stated seed.

Dataset	Model	C0 control	C0 aware	C1-C3 control	C1-C3 aware	Δ transformed F1	LRG reduction
FNIR	Logistic regression	1.000	1.000	0.429	0.534	+ 0.105	+ 0.105
FNIR	Linear SVM	1.000	1.000	0.594	0.441	-0.152	-0.152
FNIR	Multinomial NB	1.000	1.000	0.834	0.820	-0.014	-0.014
FNIR	Random forest	1.000	1.000	0.523	0.554	+ 0.031	+ 0.031
FNIR	1D-CNN	1.000	1.000	0.412	0.432	+ 0.020	+ 0.020
FNIR	LSTM	1.000	1.000	0.374	0.460	+ 0.086	+ 0.086
FNIR	BiLSTM	1.000	1.000	0.386	0.532	+ 0.147	+ 0.147
FNIR	CNN-LSTM	1.000	1.000	0.470	0.526	+ 0.056	+ 0.056
FNIR	CNN-BiLSTM	1.000	1.000	0.425	0.466	+ 0.041	+ 0.041
Monkeypox	Logistic regression	0.940	0.940	0.879	0.953	+ 0.075	+ 0.075
Monkeypox	Linear SVM	0.980	0.940	0.858	0.967	+ 0.109	+ 0.149
Monkeypox	Multinomial NB	0.940	0.960	0.762	0.980	+ 0.218	+ 0.198
Monkeypox	Random forest	0.940	0.920	0.846	0.933	+ 0.087	+ 0.107
Monkeypox	1D-CNN	0.919	0.940	0.781	0.872	+ 0.091	+ 0.070
Monkeypox	LSTM	0.879	0.857	0.683	0.933	+ 0.250	+ 0.272
Monkeypox	BiLSTM	0.839	0.940	0.784	0.913	+ 0.128	+ 0.027
Monkeypox	CNN-LSTM	0.900	0.920	0.709	0.913	+ 0.204	+ 0.184
Monkeypox	CNN-BiLSTM	0.940	0.900	0.765	0.933	+ 0.168	+ 0.208

Cross-dataset results in Table 13 show that transformation-aware augmentation did not produce domain-general robustness. Across the six evaluated transfer models in the FNIR-to-Monkeypox direction, the mean transformed macro-F1 changed only from 0.407 to 0.412 (a mean gain of + 0.005). In

the reverse Monkeypox-to-FNIR direction, mean transformed macro-F1 decreased from 0.565 to 0.542 (mean gain − 0.023). The individual results were likewise mixed: FNIR-to-Monkeypox gains for logistic regression (+ 0.056), Linear SVM (+ 0.029) and Multinomial Naive Bayes (+ 0.008) were offset by weak absolute transfer performance and negative BiLSTM/CNN-LSTM changes; Monkeypox-to-FNIR improved Random Forest (+ 0.062) and Multinomial Naive Bayes (+ 0.021) but reduced logistic regression (-0.056), Linear SVM (-0.055), BiLSTM (-0.023) and CNN-LSTM (-0.090). RQ5 therefore supports in-domain mitigation, especially on Monkeypox, but not cross-dataset generalisation.

Table 13
Cross-dataset RQ5 sensitivity analysis for four classical models and the selected BiLSTM and CNN-LSTM neural models.

Direction	Model	C1-C3 control	C1-C3 aware	Δ transformed F1
FNIR→Monkeypox	Linear SVM	0.415	0.444	+ 0.029
FNIR→Monkeypox	Logistic regression	0.333	0.390	+ 0.056
FNIR→Monkeypox	Multinomial NB	0.630	0.638	+ 0.008
FNIR→Monkeypox	Random forest	0.333	0.333	+ 0.000
Monkeypox→FNIR	Linear SVM	0.579	0.524	-0.055
Monkeypox→FNIR	Logistic regression	0.628	0.573	-0.056
Monkeypox→FNIR	Multinomial NB	0.662	0.684	+ 0.021
Monkeypox→FNIR	Random forest	0.520	0.582	+ 0.062
FNIR→Monkeypox	BiLSTM	0.359	0.333	-0.026
FNIR→Monkeypox	CNN-LSTM	0.372	0.333	-0.039
Monkeypox→FNIR	BiLSTM	0.453	0.430	-0.023
Monkeypox→FNIR	CNN-LSTM	0.548	0.458	-0.090

6. Scope of the Empirical Evaluation

The empirical C1–C3 intervention and RQ5 mitigation analyses reported in Section 5 comprise four classical classifiers and five task-trained neural/hybrid architectures. All reported performance metrics, statistical comparisons and mitigation findings are restricted to these nine executed models. Published findings from other studies are discussed only as contextual evidence and are not treated as results of the present experiments.

The empirical conclusions are restricted to the nine executed models, the two analysed health corpora and the ChatGPT GPT-5.6 Sol transformation regime. External replication should investigate whether the observed C0–C3 effects and mitigation patterns generalise across additional detector architectures, independently generated transformation sets and different Generative-AI model families.

7. Discussion

7.1 Principal findings and scientific contribution

The findings advance the evaluation of health-misinformation detection in five connected ways. First, high original-text classification performance did not establish robustness to Generative-AI reformulation. Second, transformation-induced errors differed by dataset, with false-positive inflation dominating in COVID-19 FNIR and misinformation false negatives increasing substantially in mpox. Third, neutral, ELM central-route and ELM peripheral-route transformations produced distinguishable linguistic changes, although no condition was consistently the most damaging (Tables 9–10; Fig. 4). Fourth, structural compression and increased readability grade level were stronger pooled correlates of detector vulnerability than the other persuasion-related features examined (Table 11). Fifth, transformation-aware training improved within-dataset robustness, particularly in mpox, but benefits were inconsistent in FNIR and did not reliably transfer across datasets (Tables 12–13; Fig. 5).

These findings contribute to a broader shift from static benchmark classification toward behavioural robustness. Systematic reviews of health-misinformation detection document increasingly strong machine-learning and transformer results, but also substantial heterogeneity in datasets and evaluation practices [13, 14]. The present study adds an intervention that changes message presentation while holding the source root and label fixed. The resulting performance loss therefore reveals a property that ordinary random-split evaluation cannot measure.

7.2 Near-ceiling C0 performance is not evidence of transformation invariance

Table 5 confirms that the intervention was applied to competent measurement instruments rather than weak baselines. FNIR macro-F1 on the full held-out C0 test partition ranged from 0.991 to 0.999 across the executed models, and Monkeypox baselines were also strong. Those results are consistent with previous pandemic misinformation work in which conventional, neural and transformer systems achieve high in-domain discrimination [19, 20, 21]. The C1-C3 results show why those numbers should not be interpreted as evidence of linguistic invariance.

This distinction mirrors the motivation of behavioural testing in NLP. [36] argue that held-out accuracy can overestimate reliability because it averages over naturally occurring examples rather than explicitly testing model capabilities. Semantically constrained perturbation work likewise shows that small or meaning-preserving changes can expose brittle decision rules [31, 32]. In the misinformation setting, [30], [1] and [2] independently show that LLM generation or paraphrasing changes detector behaviour. The present health-specific results therefore support a stricter definition of detector competence: a system should be evaluated not only on unseen texts, but also on alternative linguistic realisations of the same underlying claims.

The contrast between FNIR and mpox is central to interpretation. In FNIR, transformation-induced false-positive inflation occurred despite relatively strong residual ranking discrimination. This suggests that the C0 operating threshold became poorly aligned with the transformed-text distribution, consistent with sensitivity to linguistic register and dataset-specific cues rather than straightforward misinformation evasion. The distinction is important because ranking discrimination and operational classification performance may diverge substantially under distribution shift.

FNIR is particularly susceptible to this interpretation because its reliable and misinformation classes originate from different source regimes. [37] show that source overlap and collection procedures can create shortcuts that support high apparent accuracy without robust task learning, and [3] identify register variation as a confound in fake-news analysis. The current experiment provides a direct perturbational demonstration of that concern: when the surface register of both classes is normalised by Generative AI, a decision boundary that was almost perfect on C0 becomes operationally unstable. [43] provide the relevant calibration perspective, namely that strong discrimination does not guarantee that confidence estimates or decision thresholds remain reliable under changed conditions.

Monkeypox shows a qualitatively different risk. Its mean misinformation FNR increased from 0.076 at C0 to 0.409, 0.276 and 0.344 under C1-C3. Here, rewritten misinformation itself is more often classified as informational or reliable. The operational consequence is therefore missed misinformation rather than primarily excessive false alarms. In health-information moderation, those errors have different implications: false positives can suppress or delay legitimate information, while false negatives allow misleading health claims to pass undetected. Reporting only accuracy or macro-F1 would obscure this distinction, which is why the separate FPR and FNR results are integral to the analysis.

7.3 Transformation type changes message form, but does not create a universal severity ordering

RQ2 rejects the simple assumption that a single prompt condition is always the strongest attack. Across the combined model-dataset panel, the C1-C3 omnibus difference was significant, but pooled pairwise contrasts did not remain significant after Holm adjustment. FNIR showed a condition effect, including lower macro-F1 under C2 than C3 after correction, whereas Monkeypox did not show a reliable omnibus difference. The effect of transformation type is therefore conditional on the corpus and detector, rather than on an intrinsic ranking of neutral, central-route, and peripheral-route rewriting.

The feature analysis explains part of that heterogeneity. C2 produced the largest mean increase in FKGL and evidence cues, which is consistent with the intended emphasis on explanatory organisation. C3 produced the strongest increases in certainty, urgency and normative/social cues, consistent with the intended peripheral emphasis. C1 primarily shortened and regularised prose, but it also changed evidence and sentence-structure measures. The conditions therefore induced different average profiles without being linguistically orthogonal. This is important theoretically. ELM distinguishes central and peripheral routes at the level of persuasion processes [23], but a generated text can simultaneously change complexity, evidence organisation, certainty and salience. [25] and [22] demonstrate the value of

ELM-informed features for health-misinformation modelling; the present findings indicate that these features should be treated as overlapping computational descriptors rather than as direct measures of cognitive processing.

7.4 Observed vulnerability mechanisms are structural and context-dependent

The RQ4 models narrow the mechanistic claim to variables actually supported by the data. A positive change in word count was protective against a new error (OR = 0.412, 95% CI 0.304–0.556), indicating that stronger compression was associated with greater vulnerability. Increased FKGL raised the odds of a new error (OR = 1.643, 95% CI 1.133–2.382), and both variables also predicted confidence loss. In contrast, evidence, certainty, urgency and normative/social deltas were not independently significant in the pooled adjusted new-error model. A feature can therefore be strongly manipulated by the generator without independently explaining model failure once correlated changes, condition, detector and dataset are taken into account.

The interaction tests reinforce this caution. Word-count and type-token-ratio effects differed by dataset and detector, while certainty and normative/social shifts also showed dataset-dependent associations. These are computational associations with a learned decision surface, not evidence that the corresponding psychological construct caused persuasion or misinformation acceptance. The finding also complements [1], who identified sentiment shift as one explanation for laundering failures in their setting. In the present health-domain analysis, structural compression and readability were stronger pooled correlates for the retained probabilistic detectors. The difference illustrates why vulnerability mechanisms cannot be assumed to transfer unchanged across corpora, generators or model families.

7.5 Generative AI is both a source of detector stress and a defensive training resource

The study supports a dual-use interpretation of Generative AI. [27] demonstrate that LLMs can generate convincing disinformation, while [28] and [29] emphasise that the same technology can support detection and mitigation. In the present experiment, Generative AI first serves as the stressor, producing claim-preserving C1-C3 variants that expose weaknesses in detectors trained on C0. It then acts defensively by supplying transformation-aware augmentation for RQ5. This is not evidence that Generative AI directly reduces the prevalence of misinformation or its diffusion on social media, because those outcomes were not measured. It is evidence that generated variants can be used to harden detectors against linguistic reformulation.

The Monkeypox mitigation result is the strongest support for that defensive use. Transformation-aware training increased mean C1-C3 macro-F1 by 0.148 across all nine evaluated models; all nine improved, and the model-panel Wilcoxon test was significant ($p = 0.0039$). Mean LRG fell by 0.143 while panel-level C0 macro-F1 changed by only + 0.004. The improvement spans the linear, probabilistic, and neural families, so it is not attributable to a single architecture. This pattern is consistent with the adversarial

training literature showing that exposure to constrained variants can reduce perturbation sensitivity [31, 32, 35]. It also supports the LLM-era recommendation that detectors should be trained with realistic machine-generated or machine-paraphrased content rather than assuming a purely human-written deployment distribution [30].

FNIR prevents an overly broad mitigation claim. Its mean transformed-text gain was only 0.035, seven of nine models improved, and the panel-level shift was not statistically significant ($p = 0.164$). Linear SVM deteriorated despite improvement in several other models. The defensive value of Generative-AI augmentation is therefore empirical and model-specific, not automatic. This aligns with the broader proactive-defence literature, which emphasises robustness as a system property that requires evaluation across multiple threats and distributions rather than a single augmentation recipe [45].

The cross-dataset component of RQ5 draws the clearest boundary around the mitigation claim. Across the evaluated transfer models, FNIR-to-Monkeypox transformation-aware training increased the mean transformed macro-F1 by only + 0.005, while Monkeypox-to-FNIR decreased it by -0.023. Individual models moved in both directions and absolute cross-dataset performance remained weak. Training on transformed examples therefore helped models learn useful invariances within a corpus without reliably producing a domain-independent representation of veracity.

This result is consistent with an extensive generalisation literature. [40] showed that fake-news detection degrades across domains; [37] demonstrated that dataset artefacts can inflate within-dataset results; [38] linked cross-dataset performance to dataset bias and annotation granularity; and [41] showed that misinformation models must generalise across multiple axes including time, topic, event and publisher. [42] similarly found that model comparisons change when evaluation moves from in-domain to cross-domain settings. The present RQ5 result extends that evidence to Generative-AI transformation: robustness to unseen rewrites of in-domain roots is not equivalent to robustness to a different health corpus.

The executed models do not form a single robustness hierarchy. On FNIR, Multinomial Naive Bayes had the smallest mean LRG (0.200), whereas several recurrent and hybrid models exceeded 0.62. On Monkeypox, Random Forest had the smallest mean LRG (0.060), while LSTM reached 0.388 and CNN-LSTM changed sharply across conditions. These results show that greater C0 complexity does not guarantee greater robustness to transformation. Sparse lexical, recurrent and nonlinear representations can depend on different aspects of surface form, so their failure modes should be measured rather than inferred from architecture labels.

7.6 Implications for health-misinformation detection and deployment

The first practical implication is that health-misinformation detection systems should be evaluated along at least two complementary dimensions: original-distribution discrimination and transformation robustness. A detector that performs exceptionally well on unmodified content (C0) but exhibits

substantial performance deterioration following semantically faithful reformulation cannot be considered robust solely on the basis of its conventional classification metrics. This distinction is particularly important in health-information environments, where the incorrect identification of misinformation and the suppression of reliable information may have consequential implications for public health communication and informed healthcare decision-making [9, 11].

Second, ranking metrics, threshold-dependent performance measures and class-specific error rates should be evaluated jointly. The COVID-19 FNIR results demonstrate the importance of this distinction: although AUROC remained comparatively high following transformation, macro-F1 deteriorated substantially, and the false-positive rate exceeded 0.83 across C1–C3. This indicates that considerable ranking discrimination was retained despite substantial deterioration in classification performance at the original C0 operating threshold. Accordingly, calibration-aware evaluation is important for distinguishing changes in discriminative capability from shifts in operational decision boundaries [43]. Potential safeguards include transformation-aware calibration, uncertainty estimation and appropriately validated abstention policies. However, these approaches represent practical implications of the observed error patterns rather than interventions directly evaluated in the present study.

Third, Generative AI can be employed defensively as a mechanism for stress-testing and augmenting health-misinformation detection systems. Previously verified or fact-checked claims can be reformulated under controlled transformation conditions to determine whether classification decisions are sensitive to changes in fluency, linguistic complexity or rhetorical register. Transformations that satisfy appropriate claim-fidelity requirements may subsequently be incorporated into model training to improve robustness against linguistic reformulation. The RQ5 findings provide empirical support for this approach, particularly in the mpox dataset, where transformation-aware training improved classification performance on previously unseen source texts. Nevertheless, these improvements were dataset-dependent and did not consistently transfer across health-information domains. Furthermore, improved detector robustness should not be interpreted as direct evidence of reduced misinformation exposure, changes in health behaviour or improved clinical outcomes, as these effects were not evaluated.

Implications for personalised digital health information services

The findings are relevant to the evaluation of AI-supported information services that may contribute to personalised healthcare. Such systems may process health claims presented through neutral explanations, evidence-oriented messages and emotionally salient narratives. The observed instability across transformation conditions suggests that conventional benchmark discrimination alone may be insufficient when evaluating potential misinformation-detection components for patient-facing digital applications. Excessive false positives may restrict access to reliable health information, whereas misinformation false negatives may leave misleading claims unflagged. These distinct information-quality risks support application-specific assessment using validated decision thresholds, class-specific error measures, contextual verification and appropriate human oversight. However, the present study provides computational evidence concerning detector behaviour rather than direct evidence of clinical

effectiveness or patient safety. Establishing whether improved detector robustness translates into more reliable individualised health-information delivery, better healthcare decisions or improved patient outcomes requires subsequent application-specific validation.

Several design requirements follow directly from the evidence. Both veracity classes should be transformed symmetrically so that machine register does not become a surrogate class label, as emphasised by [2]. Source roots and all of their descendants should remain in a single partition to prevent paraphrase leakage. Operating thresholds should be fixed before the stress test, with AUROC, AUPRC, macro-F1, FPR and FNR reported together. Claim fidelity should be treated as an eligibility condition rather than inferred solely from semantic similarity, consistent with the adversarial-validity concerns raised by [34].

Robustness should also be stratified by generalisation level. The present results support three distinct questions: whether a model handles transformed versions of unseen roots from the same corpus, whether mitigation transfers to a different corpus, and whether it survives a different generator family. Evidence for one level does not establish the others. This separation is consistent with recent calls for source-disjoint, cross-domain and multidimensional out-of-distribution evaluation in misinformation research [37, 38, 41]. It also argues against benchmark designs in which a single near-ceiling random-split score is treated as evidence of real-world robustness.

The study extends theory-informed health-misinformation modelling by moving from feature inclusion to controlled feature displacement. Previous work has shown that ELM-derived readability, sentiment, and heuristic cues can improve classification [25, 22], and cross-pandemic analysis has established that health corpora differ systematically in linguistic profile [5]. The present experiment uses Generative AI to alter observable message properties while keeping the root label fixed, then measures the resulting detector response. This design makes theory-informed features useful as descriptors of transformation, not merely as additional predictors.

The results also impose an important limit on theoretical interpretation. C2 and C3 produced feature changes broadly consistent with their intended central and peripheral emphasis, but RQ4 did not show that all persuasion-related deltas independently predicted detector failure. ELM and TPB therefore organise the manipulation and feature taxonomy; they do not convert text-derived cues into evidence about human cognition or behavioural intention. That boundary keeps the fully computational study conceptually coherent and avoids claiming psychological mechanisms that were not measured.

8. Strengths and Limitations

Strengths include a paired C0-C3 intervention in which each source root serves as its own control, balanced veracity within the intervention cohort, symmetric rewriting of both classes, two health-misinformation corpora with different linguistic regimes, a nine-model executed classical/CNN-recurrent panel, frozen pre-intervention thresholds, explicit FPR/FNR analysis, a fully implemented and auditable computational feature inventory, pooled and stratified mechanism analyses, and an RQ5 mitigation

design with disjoint augmentation and mitigation-test roots. The RQ5 control receives the same additional C0 roots as the transformation-aware regime, so the comparison isolates exposure to transformed descendants rather than merely exposure to more claim families.

The principal limitation concerns transformation fidelity. Although the ChatGPT generation prompts explicitly required preservation of the underlying health claims, and structured source-descendant consistency checks were applied, independent natural language inference and external clinical-claim adjudication were not performed. Consequently, complete semantic and clinical equivalence cannot be established for every transformed item. This limitation is particularly relevant to C3, where changes in certainty-related language may alter the interpretation of a medical claim. The reported effects should therefore be understood as exploratory estimates of detector sensitivity to constrained generative-AI rewriting rather than definitive estimates based on independently verified proposition-preserving transformations.

A second limitation concerns potential semantic overlap between the original dataset partitions. Exact-text deduplication was conducted before model development, and generated descendants remained associated with their original source roots. However, semantic near-duplicate claim-family clustering was not implemented across the original training, validation and test partitions. Root-disjoint evaluation therefore does not guarantee the absence of semantically equivalent claims across partitions. This limitation warrants particular consideration when interpreting near-ceiling original-text performance on COVID-19 FNIR, where source-specific and linguistic regularities may influence classification.

Third, the intervention included 100 source roots per dataset, while the mitigation analysis used 50 untouched evaluation roots per dataset. These relatively small samples limit the precision and generalisability of the observed error rates and mitigation effects. The neural mitigation experiments were conducted using a single random seed, and model-level statistical comparisons do not capture the full uncertainty associated with sampling variation and repeated neural training. Paired source-root bootstrap confidence intervals were estimated for the four classical classifiers in the supplementary mitigation analysis. However, these intervals describe sampling uncertainty conditional on the fitted models and the fixed experimental split; they do not capture uncertainty arising from neural-model training, alternative dataset partitions or repeated Generative-AI transformations. Additional repeated-seed and independent-generation experiments are therefore required to strengthen the inferential evidence.

Fourth, the experiment used a single generative model and one transformed descendant per condition, preventing conclusions about robustness across different generator families or independently sampled rewrites. Computational mechanism analyses were restricted to the retained probabilistic outputs of logistic regression and Multinomial Naive Bayes, while the linguistic feature dictionaries represented simplified operationalisations of rhetorical and persuasion-related characteristics.

Finally, the observed error patterns and mitigation effects were dataset-dependent, and cross-dataset transfer was inconsistent. Differences between COVID-19 FNIR and mpox may reflect dataset

construction, source characteristics, annotation practices and linguistic distributions rather than disease context alone. Furthermore, although these findings are relevant to evaluating potential information-quality safeguards in personalised digital healthcare, the study did not examine patient-specific information delivery, clinical implementation, healthcare decisions or patient outcomes. The clinical relevance of transformation-aware misinformation detection therefore requires independent, application-specific validation.

9. Conclusion

This study provides direct computational evidence that Generative-AI rewriting can materially alter health-misinformation detector behaviour even when the experimental intention is to preserve the underlying proposition. Across 200 balanced source roots, FNIR mean macro-F1 fell from 0.987 at C0 to 0.428–0.479 under C1-C3, while Monkeypox fell from 0.922 to 0.731–0.781.

The failure modes were different: FNIR was dominated by false-positive inflation and decision-boundary/register shift, whereas Monkeypox showed substantial increases in misinformation false negatives. Taken together, these findings show that health-misinformation detection is not adequately characterised by static benchmark classification alone; robustness to Generative-AI reformulation is a distinct property that must be measured explicitly.

The transformations also produced measurable, condition-specific feature changes, and pooled mechanism models associated stronger text compression and increased FKGL with greater vulnerability. Transformation-aware training substantially improved unseen-root robustness on Monkeypox (mean transformed-text gain + 0.148; $p = 0.0039$) but produced smaller, heterogeneous FNIR gains (+ 0.035; $p = 0.1641$), and cross-dataset transfer did not improve reliably.

Generative AI consequently occupies a dual role: it can expose linguistic vulnerabilities that destabilise existing detectors, but the same transformation capability can be repurposed defensively to generate challenging training examples and improve resilience to previously unseen formulations. The evidence therefore supports transformation-aware augmentation as an in-domain robustness intervention, not as a complete solution to domain shift.

Ultimately, as generative AI tools continue to lower the barrier for laundering misleading content, static accuracy metrics can no longer serve as the sole validation standard for health-related text classifiers. Robust health-misinformation detection must be evaluated under symmetric register control, multiple transformation conditions, explicit error-type reporting, root-disjoint mitigation tests and external replication across datasets, generators and model families. These findings support the inclusion of symmetric transformation testing, class-specific error analysis, root-disjoint evaluation and external replication when assessing automated health-information classifiers. Such evaluation may inform the development of more reliable information-quality safeguards for personalised digital health applications. However, independent semantic-fidelity validation, stronger claim-family separation, repeated-seed

experiments and application-specific clinical evaluation remain necessary before conclusions can be drawn about deployment readiness or patient-level benefits.

Declarations

Declaration of generative artificial intelligence use

OpenAI ChatGPT (GPT-5.6 Sol) was used both to generate the C1–C3 experimental text transformations described in the Methods and to assist with manuscript organisation, language refinement, and identifying and debugging Python programming errors. AI-generated suggestions were assessed by the author against the underlying research data, computational outputs, methodological specifications and relevant scholarly literature. The author reviewed the scientific content, checked numerical results against analytical outputs, and retained responsibility for selecting, implementing and validating the reported methods. AI assistance was not treated as independent empirical evidence or verification of the study's conclusions. The author takes full responsibility for the accuracy, integrity and final content of the manuscript.

Ethics approval and consent to participate

This study involved secondary computational analysis of previously published health-misinformation datasets and did not involve recruitment of human participants or new interactions with social-media users. The analyses used existing text, supplied veracity labels and derived computational features. No attempt was made to re-identify individual users; findings are reported primarily in aggregate form, and unnecessary republication of user identifiers and profile information was avoided. The original datasets were accessed subject to their applicable terms. No prospective participant consent was sought because no new participant interactions occurred. The study did not independently adjudicate the medical accuracy of individual claims or establish complete semantic equivalence between source and rewritten texts.

Consent for publication

Not applicable. The manuscript does not reproduce identifiable individual-level information, photographs or case histories.

Availability of data and materials

The study used previously published health-misinformation datasets. COVID-19_FNIR is available through IEEE DataPort (https://doi.org/10.21227/b5bt-5244), and the Monkeypox Twitter dataset is available through Kaggle (https://doi.org/10.34740/KAGGLE/DS/2352848). The original datasets were not generated by the author and should be obtained directly from their providers in accordance with applicable access and reuse conditions. The frozen C0–C3 intervention texts are not redistributed publicly; complete reproduction of text-dependent experiments therefore requires separate access to these records.

Code availability

Python source code, computational workflows, configuration files, frozen aggregate experimental results, the intervention-text integrity manifest and supplementary classical-model mitigation analyses are archived on Zenodo (https://doi.org/10.5281/zenodo.23645855). The archive includes documentation and scripts for data preprocessing, classifier training, transformation-robustness evaluation, statistical analysis and paired source-root bootstrap estimation. These materials support verification of the reported analyses, subject to the data-access limitations stated above.

Competing interests

The author declares that he has no competing interests.

Funding

This research received no external funding.

Authors' contributions

MS was the sole author and was responsible for conceptualisation, study design, data preparation, computational implementation, Generative-AI transformation procedures, statistical analysis, interpretation of findings, manuscript drafting and revision. The author read and approved the final manuscript.

Acknowledgements

Not applicable.

Clinical trial registration

Clinical trial number: not applicable.

References

1. Das RK, Dodge J. Fake news detection after LLM laundering: Measurement and explanation. arXiv:2501.18649. 2025.
2. Mehdiyev J, Aliguliyev R. Cross-LLM paraphrase laundering: A register-controlled evaluation of fake news detectors. AI. 2026;7(8):297. https://doi.org/10.3390/ai7080297
3. Baissa B, Fuoli M, Grieve J. Veracity and register in fake news analysis. Linguistics Vanguard. 2025. https://doi.org/10.1515/lingvan-2024-0069
4. Saenz JA, Gopal SRK, Shukla D. COVID-19 Fake News Infodemic Research Dataset (COVID19-FNIR Dataset). IEEE DataPort. 2021.

5. Sikosana M, Maudsley-Barton S, Ajao O. Linguistic patterns in pandemic-related content: A comparative analysis of COVID-19, Constraint, and Monkeypox datasets. Frontiers in Artificial Intelligence. 2025;8:1627522. https://doi.org/10.3389/frai.2025.1627522
6. World Health Organization. Global strategy on digital health 2020–2027. Geneva: World Health Organization. 2025. https://www.who.int/publications/i/item/9789240116870
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-596. https://doi.org/10.1001/jamainternmed.2023.1838
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887. https://doi.org/10.3390/healthcare11060887
9. Chou WYS, Oh A, Klein WMP. Addressing health-related misinformation on social media. JAMA. 2018;320(23):2417-2418. https://doi.org/10.1001/jama.2018.16865
10. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization. 2021. https://www.who.int/publications/b/58847
11. Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine. 2019;240:112552. https://doi.org/10.1016/j.socscimed.2019.112552
12. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res. 2021;23(1):e17187. https://doi.org/10.2196/17187
13. Sanaullah AR, Das A, Das A, Kabir MA, Shu K. Applications of machine learning for COVID-19 misinformation: A systematic review. Social Network Analysis and Mining. 2022;12:94. https://doi.org/10.1007/s13278-022-00921-9
14. Schlicht IB, Fernandez E, Chulvi B, Rosso P. Automatic detection of health misinformation: A systematic review. Journal of Ambient Intelligence and Humanized Computing. 2024;15:2009-2021. https://doi.org/10.1007/s12652-023-04619-4
15. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter. 2017;19(1):22-36. https://doi.org/10.1145/3137597.3137600
16. Kim Y. Convolutional neural networks for sentence classification. Proceedings of EMNLP 2014, 1746-1751. 2014. https://doi.org/10.3115/v1/D14-1181
17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
18. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing. 1997;45(11):2673-2681. https://doi.org/10.1109/78.650093
19. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, Ekbal A, Das A, Chakraborty T. Fighting an infodemic: COVID-19 fake news dataset. In Combating Online Hostile Posts in Regional Languages during Emergency Situation, 21-29. Springer. 2021. https://doi.org/10.1007/978-3-030-73696-5_3

20. Alghamdi J, Lin Y, Luo S. Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems. 2023;274:110642. https://doi.org/10.1016/j.knosys.2023.110642
21. Sikosana M, Ajao O, Maudsley-Barton S. A comparative study of hybrid models in health misinformation text classification. Proceedings of the 4th International Workshop on Open Challenges in Online Social Networks, 18-25. 2024. https://doi.org/10.1145/3677117.3685007
22. Sikosana M, Maudsley-Barton S, Ajao O. Advanced health misinformation detection through hybrid CNN-LSTM models informed by the Elaboration Likelihood Model (ELM). 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications, 1-11. 2025. https://doi.org/10.1109/ACDSA65407.2025.11166406
23. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. Advances in Experimental Social Psychology. 1986;19:123-205. https://doi.org/10.1016/S0065-2601(08)60214-2
24. Ajzen I. The theory of planned behavior. Organizational Behavior and Human Decision Processes. 1991;50(2):179-211. https://doi.org/10.1016/0749-5978(91)90020-T
25. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management. 2021;58(1):102390. https://doi.org/10.1016/j.ipm.2020.102390
26. Sikosana M, Maudsley-Barton S, Ajao O. A multi-branch feature fusion approach for health misinformation detection and propagation. arXiv:2609.00403. 2026.
27. Vykopal I, Pikuliak M, Srba I, Moro R, Macko D, Bielikova M. Disinformation capabilities of large language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 14830-14847. 2024. https://doi.org/10.18653/v1/2024.acl-long.793
28. Chen C, Shu K. Combating misinformation in the age of LLMs: Opportunities and challenges. AI Magazine. 2024;45(3):354-368. https://doi.org/10.1002/aaai.12188
29. Park S, Nan X. Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. AI & Society. 2026;41:1501-1515. https://doi.org/10.1007/s00146-025-02620-3
30. Su J, Cardie C, Nakov P. Adapting fake news detection to the era of large language models. Findings of the Association for Computational Linguistics: NAACL 2024, 1473-1490. 2024. https://doi.org/10.18653/v1/2024.findings-naacl.95
31. Ribeiro MT, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 856-865. 2018. https://doi.org/10.18653/v1/P18-1079
32. Iyyer M, Wieting J, Gimpel K, Zettlemoyer L. Adversarial example generation with syntactically controlled paraphrase networks. Proceedings of NAACL-HLT 2018, 1875-1885. 2018. https://doi.org/10.18653/v1/N18-1170
33. Garg S, Ramakrishnan G. BAE: BERT-based adversarial examples for text classification. Proceedings of EMNLP 2020, 6174-6181. 2020. https://doi.org/10.18653/v1/2020.emnlp-main.498

34. Morris J, Lifland E, Lanchantin J, Ji Y, Qi Y. Reevaluating adversarial examples in natural language. Findings of the Association for Computational Linguistics: EMNLP 2020, 3829-3839. 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.341
35. Morris JX, Lifland E, Yoo JY, Grigsby J, Jin D, Qi Y. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 119-126. 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.16
36. Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: Behavioral testing of NLP models with CheckList. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4902-4912. 2020. https://doi.org/10.18653/v1/2020.acl-main.442
37. Zhou X, Elfardy H, Christodoulopoulos C, Butler T, Bansal M. Hidden biases in unreliable news detection datasets. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2482-2492. 2021. https://doi.org/10.18653/v1/2021.eacl-main.211
38. Kishi Y, Arima Y, Iyatomi H. Fake news detection strategies under dataset bias: Using large-scale coarse-grained labels. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, Student Research Workshop, 612-621. 2026. https://doi.org/10.18653/v1/2026.eacl-srw.47
39. Calvo H, Laureano MH. High accuracy, low generalization: Structural homogeneity and cross-dataset evaluation in fake-news benchmarks. Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, 25-33. 2026. https://doi.org/10.63317/4ka6fv4hbw9z
40. Silva A, Luo L, Karunasekera S, Leckie C. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35(1):557-565. https://doi.org/10.1609/aaai.v35i1.16134
41. Verhoeven I, Mishra P, Shutova E. Yesterday's news: Benchmarking multi-dimensional out-of-distribution generalization of misinformation detection models. Computational Linguistics. 2026;52(2):619-675. https://doi.org/10.1162/coli.a.585
42. Dell'Oglio P, Bondielli A, Marcelloni F, Passaro LC. An experimental comparison of the most popular approaches to fake news detection. Inf Sci. 2026;745:123407. https://doi.org/10.1016/j.ins.2026.123407
43. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. 2017;70:1321-1330.
44. Wei J, Zou K. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of EMNLP-IJCNLP 2019, 6382-6388. 2019. https://doi.org/10.18653/v1/D19-1670
45. Liu S, Liu H, Liu A, Duan B, Zheng Q, Yan Y, Geng H, Jiang P, Liu J, Hu X. A survey on proactive defense strategies against misinformation in large language models. Findings of the Association for Computational Linguistics: ACL 2025, 18144-18155. 2025. https://doi.org/10.18653/v1/2025.findings-acl.933

46. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health. 2019;5:2055207619871808. https://doi.org/10.1177/2055207619871808
47. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015;351:h3480. https://doi.org/10.1136/bmj.h3480
48. Castle-Clarke S, Imison C. The digital patient: transforming primary care? London: Nuffield Trust. 2016.
49. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res. 2019;21(4):e12887. https://doi.org/10.2196/12887
50. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva: World Health Organization. WHO guidance announcement. 2024.
51. Sikosana M. From False to Fluent_ A Claim-Disjoint Baseline Study and Preregistered Framework for Evaluating Generative-AI Transformation of Health Misinformation [Computer software]. Zenodo. 2026. https://doi.org/10.5281/zenodo.23645855

Figures

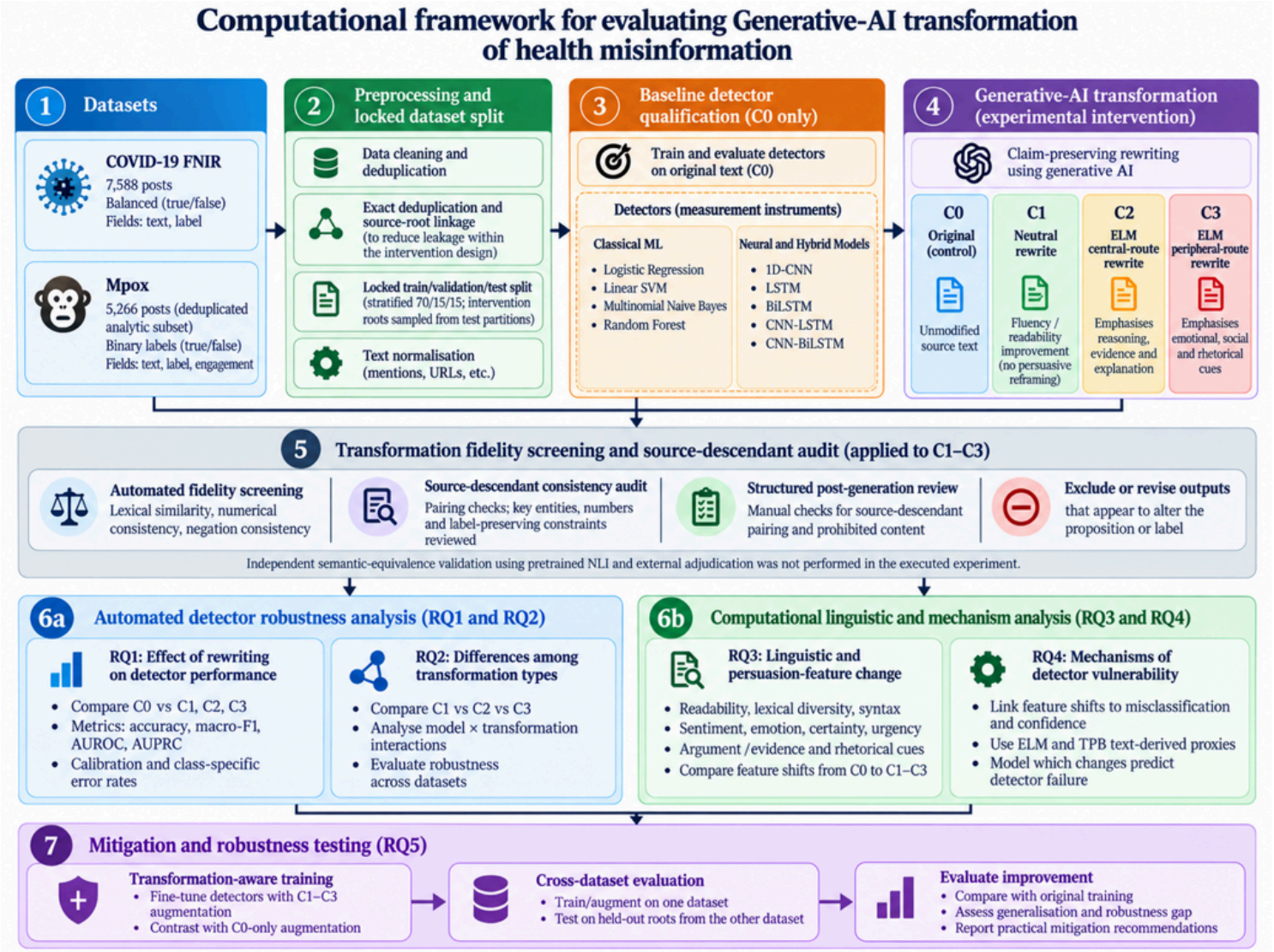

Figure 1

Fully computational study architecture. *Generative-AI transformation is the experimental intervention, while detector performance, computational linguistic-feature shifts, vulnerability mechanisms and mitigation are the measured outcomes. The implemented detector panel comprises four classical classifiers and five task-trained neural architectures. Transformation fidelity was assessed through source-descendant consistency checks and automated lexical, numerical and negation screening; independent semantic-equivalence validation using pretrained natural language inference and external adjudication was not performed. The mitigation experiment used held-out source roots and evaluated within-dataset robustness and cross-dataset transfer. Cross-generator evaluation was not performed.*

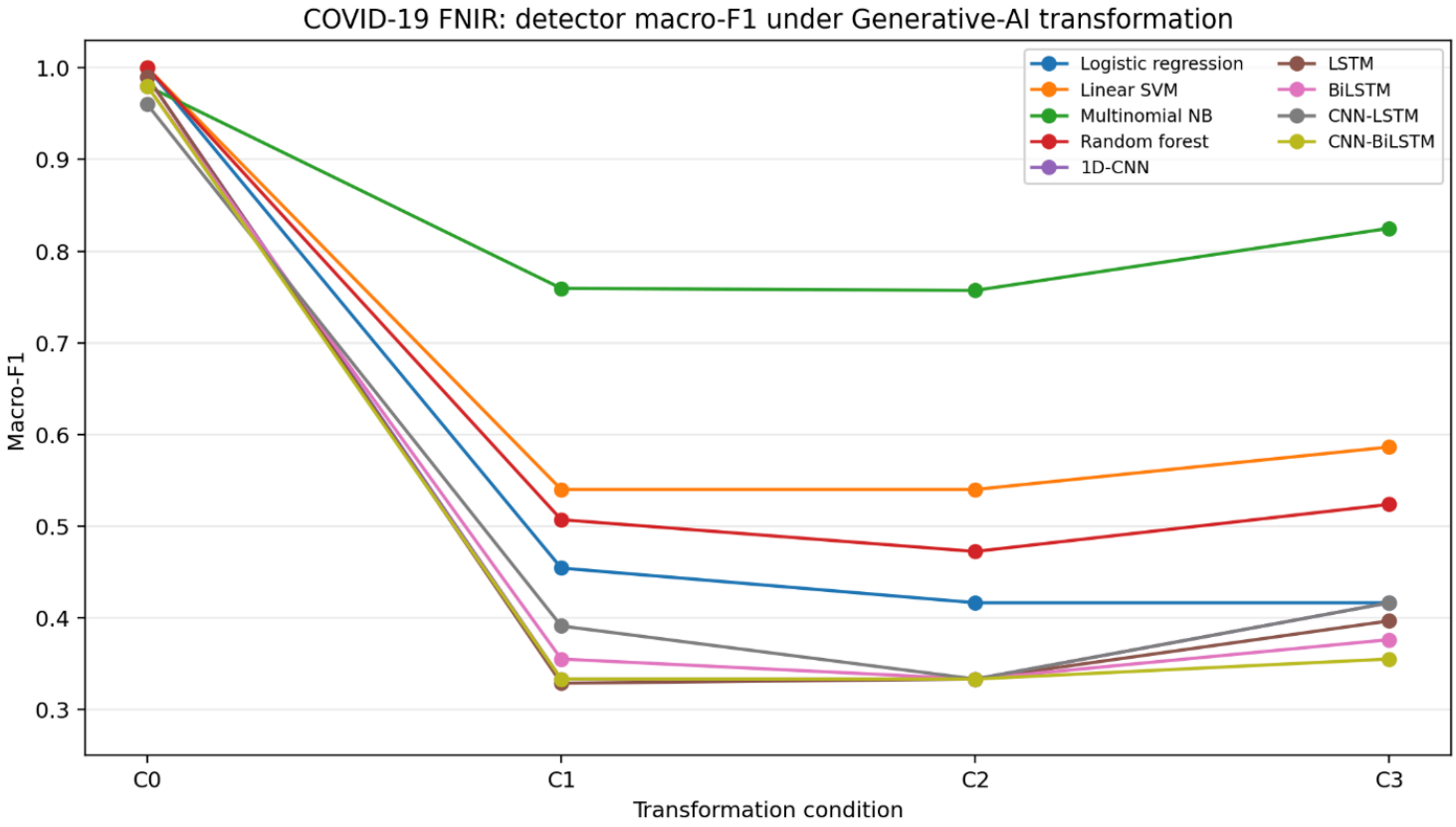

Figure 2

COVID-19 FNIR macro-F1 trajectories from original C0 text to ChatGPT transformation conditions C1-C3.

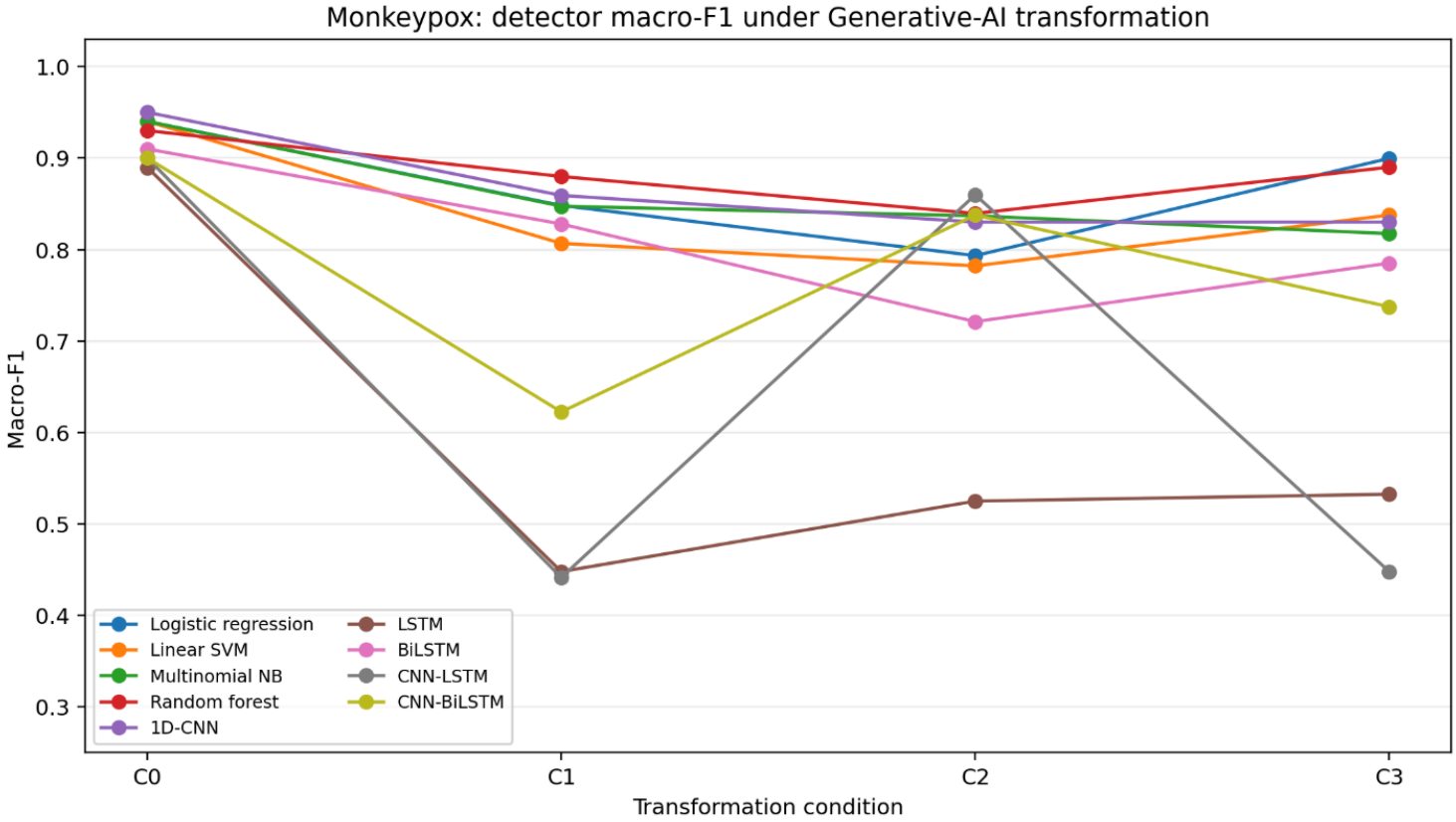

Figure 3

Monkeypox macro-F1 trajectories from original C0 text to ChatGPT transformation conditions C1-C3.

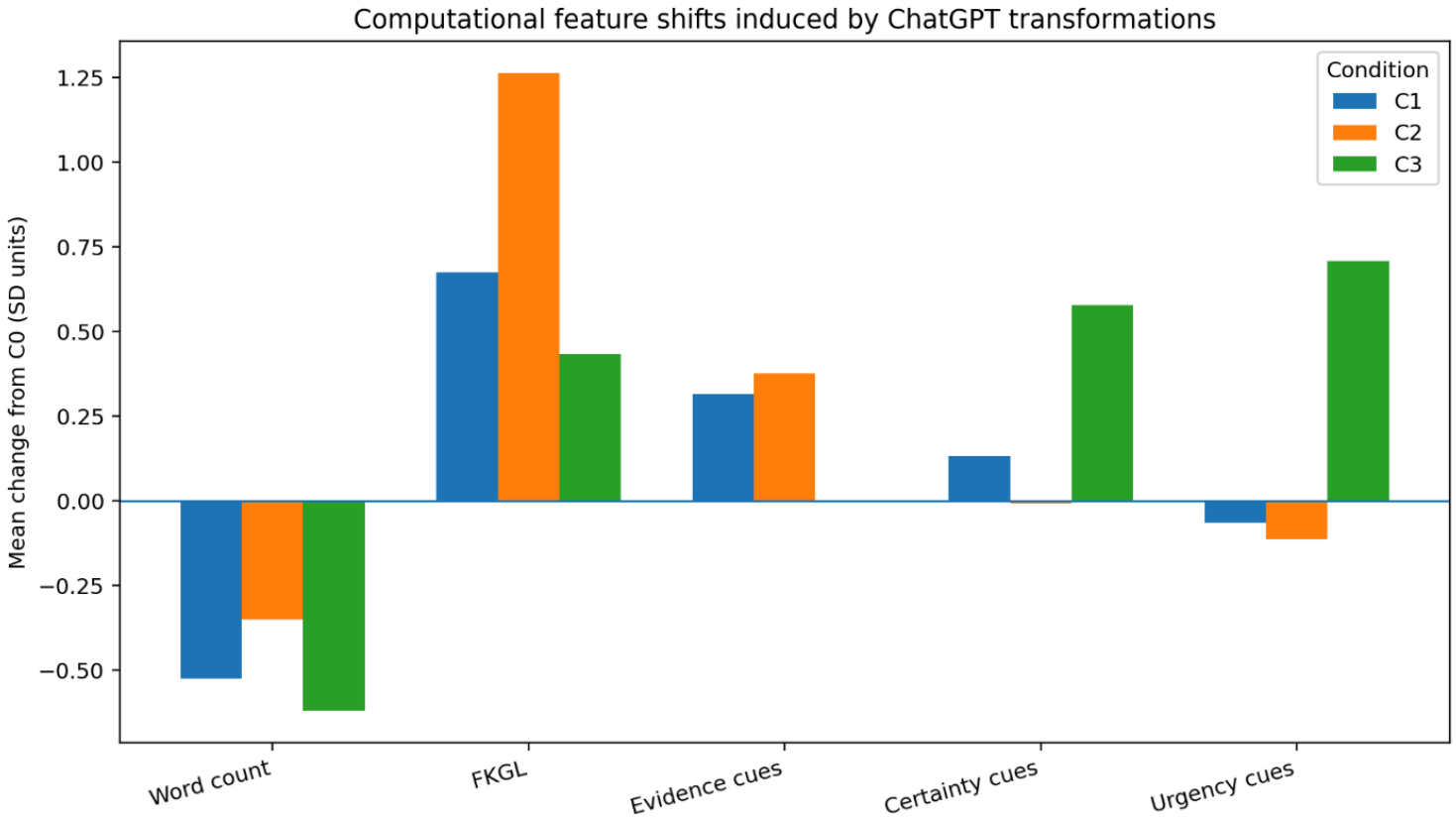

Figure 4

Standardised computational feature shifts induced by C1-C3 relative to C0.

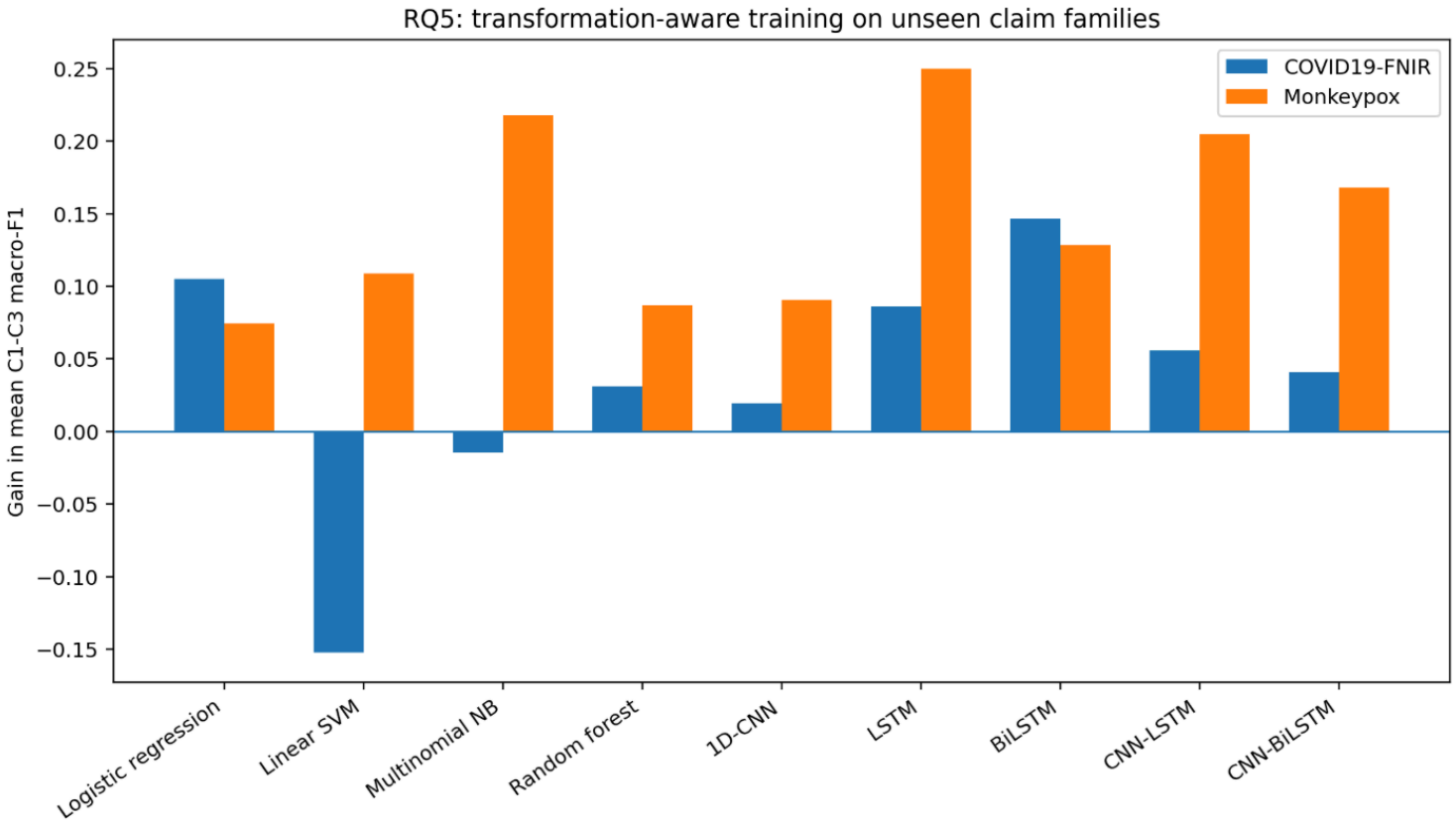

Figure 5

RQ5 gain in mean transformed-text (C1-C3) macro-F1 after transformation-aware training. Values above zero indicate improved robustness on held-out source roots.