

Article

Cross-LLM Paraphrase Laundering: A Register-Controlled Evaluation of Fake News Detectors

Jalal Mehdiyev ^{1,2,*}  and Ramiz Aliguliyev ^{1,2} 

¹ Institute of Information Technology, B. Vahabzade Str 9A, Baku AZ1141, Azerbaijan; r.aliguliyev@gmail.com

² Institute of Data Science and Artificial Intelligence, Azerbaijan Technical University, H. Javid Ave 25, Baku AZ1073, Azerbaijan

* Correspondence: jalal.mehdiyev.s@gmail.com

Abstract

Detectors built on transformer language models report near-perfect accuracy on standard fake news benchmarks, which suggests the task is almost solved. We argue that much of this accuracy reflects a confound between writing register and veracity: in common benchmarks, the real class is human-written while the fake class is machine-generated or machine-rewritten, so a detector can separate the classes by recognizing AI writing style rather than by judging truth. To evaluate this, we designed a two-regime evaluation. Phase 1 is the laundering regime, comparing untouched human-real articles against laundered fake articles. Phase 2 is register-controlled: the real class is passed through the same cross-LLM laundering chains as the fake class, so both classes are read in one machine register, and the register cue is no longer available to the detector. We train seven detectors on three datasets and evaluate each frozen detector under both regimes. Under Phase 1, detectors appear robust; under Phase 2, detection on WELFake collapses from about 99% to about 62% AUROC and the largest models approach chance. The effect is benchmark dependent, large on WELFake, mild on IFND and near zero on GossipCop, and it is confirmed by bootstrap testing with false discovery rate control. We recommend register-controlled evaluation as standard reporting practice.

Keywords: fake news detection; large language models; laundering; AI-style shift vulnerability; generative AI

1. Introduction

The spread of fake news is one of the central problems of the modern information environment. A large body of work has tried to answer it with automatic detection of fake news. The dominant approach today fine tunes a pretrained transformer language model on a labelled corpus of real and fake articles [1]. On the common benchmarks, those detectors report accuracy and AUROC values that are close to the ceiling, frequently above 98%. FakeBERT reaches 98.90% on a Kaggle corpus [2], and the hybrid model of Nasir et al. reports about 99% on ISOT [3]. That same model falls to about 0.60 on FA-KES [3], which shows that the high scores are tied to the benchmark and not only to the task. Taken at face value, such numbers suggest that detection of fake news is a solved problem. We do not read them this way. A single accuracy number is a weak frame for this task, because it lets the shortcut a detector takes stay hidden [4,5].

We do not believe that conclusion, and this paper explains why. A detector that scores 99% on a held-out test split has learned some function that separates the real articles from



Academic Editor: Ermelinda Oro

Received: 11 June 2026

Revised: 30 July 2026

Accepted: 1 August 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

the fake ones. The open question is what that function relies on. If it relies on the truth of the claims, the score is meaningful. If it touches on a surface property that happens to correlate with the label inside the benchmark, the score is an artifact of how the benchmark was built. This distinction is not academic. A detector that has learned a surface shortcut will keep its high benchmark score and still fail on any input where the shortcut and the label come apart.

Large language models made one such input cheap and common. Anyone can take a piece of fake news and ask a model to rewrite it, which produces a fluent version that carries the same false claims in new wording. We call the repeated application of those rewrites paraphrase laundering. Laundering is a realistic threat, because it is cheap, it needs no technical skill, and it preserves the disinformation payload while changing the text that a detector sees. The broad risks of large language models, including their use for low-cost text production at scale, are by now widely discussed [6]. Earlier work already showed that paraphrasing weakens detectors of machine-generated text [7]. Our interests are different. We use laundering not only as an attack but as an instrument that exposes what the detector learned in the first place.

The benchmarks used to train and to test fake news detectors share a quiet construction choice. The real class is collected from human-written journalism, while the fake class is, in a growing share of recent corpora, either generated or rewritten by a model, or at least heavily edited. Because of this, the label is correlated with the writing register. Human journalism and machine text differ in measurable ways, in lexical variety, in punctuation, in sentiment, and in the use of surface markers such as capitalization. A detector that learns to recognize AI writing style will therefore score remarkably high on the benchmark without ever modeling veracity. The reported accuracy is then a measurement of register separation, not of truth detection. We name this failure as the AI-style-shift vulnerability. By this, we mean the failure where a detector keys on register and style cues that a model rewrite can move, and not on the veracity of the text. This is a form of shortcut learning [4], close to what Schuster et al. [8] report for stylometry. Potthast et al. [9] give an early sign of it. Their stylometric study tells hyperpartisan from mainstream writing at about 0.78 F1 but tells fake from real only at about 0.46 F1 – evidence that style tracks register, not truth.

The confound is hard to see with the standard evaluation, which is why it has mostly stayed hidden. Standard evaluation compares untouched human-real text against fake text, so register and veracity move together and cannot be told apart. To separate them, the register difference must be removed while the veracity difference is kept. That is the core idea of this study.

This paper makes the following contributions.

- We introduce a register-controlled evaluation regime. The real class is passed through the same rewriting and laundering chains as the fake class; thus, both classes are read by the detector in one common machine register, and the register cue is removed. What remains is that whatever else still separates the two classes, which we characterize in Section 4.5.
- We show, on three datasets and seven detectors, that standard evaluation inflates the reported score. On WELFake, the inflation reaches about 37 AUROC points, which we attribute to register rather than to veracity.
- We report a capacity inversion. Once register is controlled, larger detectors collapse harder than smaller ones, and the two largest neural models reach chance level.
- We give a cross-benchmark spectrum of style reliance and a linguistic analysis of the mechanism, which connects the size of the collapse to the amount of register signal in each benchmark.
- We recommend register-controlled evaluation as a standard reporting practice.

Section 2 reviews related work and positions the study against research on shortcut learning. Section 3 presents the materials and methods, including the problem formulation, the laundering pipeline, the detectors and the two evaluation regimes. Section 4 reports the results. Section 5 analyses the linguistic mechanism. Section 6 discusses the implications, Section 7 states the limitations, and Section 8 concludes.

2. Related Work

Research on the detection of fake news has a long history. The survey by Zhou and Zafarani [10] organizes it along four angles: the false knowledge a story carries, its writing style, the way it propagates, and the credibility of its source. An earlier data-mining perspective is given by Shu et al. [11], who group detection by news content and by social context, with the content features split into knowledge-based and style-based. The modern mainstream sits in the first two of those angles and fine tunes a pretrained transformer language model on a labeled corpus. The model is usually BERT or a close relative, the bidirectional transformer of Devlin et al. [12], which is pretrained on large unlabeled text and then adapted to the detection task with a small classification head. On the common benchmarks, those detectors report very strong numbers, and on WELFake, transformer detectors routinely exceed 98% on every standard metric [2,3]. Those reported numbers have risen from one paper to the next, which is easy to read as steady progress. Our work does not dispute that the numbers are real, it asks what they measure.

The benchmarks themselves deserve a closer look, because their construction is where the confound enters. WELFake [1] was built by Verma et al., who merged four existing fake news collections into one larger corpus of about 72,000 articles, balanced between real and fake, to reduce overfitting. GossipCop is the celebrity and entertainment part of FakeNewsNet [13], the repository of Shu et al. that pairs news content with social context, where the labels come from a fact-checking website. IFND [14] is the Indian news corpus of Sharma and Garg, collected at claim-level, where the fake class was grown with an augmentation step because genuine fake items were scarce. Earlier corpora followed the same pattern. The LIAR dataset [15] of Wang gathered about 13,000 short statements from a political fact-checking site under six truth labels, and Pérez-Rosas et al. [5] built two datasets across seven news domains and studied the linguistic differences between fake and legitimate items. In every one of those corpora, the real class and the fake class are collected or produced by different routes, so the label is tied to how the text was made and not only to whether it is true. That construction choice is the one this paper turns into a measurement.

A separate line of work attacks text classifiers and asks whether a high benchmark score survives pressure. The CheckList method of Ribeiro et al. [16] evaluates a model with many targeted behavioral cases and shows that strong aggregate accuracy can hide systematic failures. For the detection of machine-generated text, paraphrasing is the attack that matters most. Krishna et al. [7] built DIPPER, a paraphrase model of 11 billion parameters, and showed that paraphrasing the output of a large language model evades watermarking and trained detectors alike, dropping the detection rate of one popular method from about 70% to under 5%. They answered with a retrieval-based defense. Sadasivan et al. [17] pushed the attack further with recursive paraphrasing, broke a wide range of detectors including watermarking and zero-shot methods, and proved a theoretical limit, namely that the accuracy of even the best possible detector is bounded by how far the distributions of human and machine text differ. Our laundering pipeline is related to those attacks, but we do not use it only to lower a score. We use it as a controlled transformation that lets us measure how much of a detector's score depends on the register.

The closest background for this paper is the study of shortcut learning. Geirhos et al. [4] gave the idea of its name and its frame. A shortcut is a decision rule that scores well on the test split of a benchmark but fails once the testing conditions change, and they argue that many failures of deep learning are of exactly this kind. Natural language processing has clear and well-documented cases. Gururangan et al. [18] showed that natural language inference datasets carry annotation artifacts, so a model that reads only the hypothesis and never the premise still classifies about 67% of one standard dataset correctly, because the crowd workers who wrote the hypotheses left lexical clues. McCoy et al. [19] found that inference models lean on shallow syntactic heuristics such as word overlap, and built a controlled test set, named HANS, on which those heuristics fail and strong models collapse. Niven and Kao [20] found that BERT reached about 77% on an argument comprehension task, close to the untrained human level, entirely by exploiting statistical cues in the data, and that on an adversarial version of the task every model falls to chance. Our study places the detection of fake news inside the same family. The AI-style-shift vulnerability is a shortcut, and the writing register is the spurious cue. The new part of our contribution is to measure the size of the shortcut and to give an evaluation that removes it.

Parallel literature detects machine-generated text directly, and it is useful here for one reason. GLTR, the tool of Gehrmann et al. [21], uses simple statistical signatures, the probability and the rank of each word and the entropy of the predicted distribution, and it raised the human detection rate of generated text from about 54% to 72%. DetectGPT, the method of Mitchell et al. [22], needs no trained detector at all and instead uses the curvature of the log probability of the model under test. Trained detectors were built as well. Solaiman et al. [23] released a detector together with the staged release of an early large model, Zellers et al. [24] built the Grover system and found that the best detector of neural fake news is itself a generator, and Guo et al. [25] assembled the HC3 corpus, which pairs human and chat-model answers, and trained detectors on it. The common message of that literature is that human text and machine text differ in stable and learnable ways, so a fake news detector trained on a corpus whose fake class is machine text inherits that separable signal as a free shortcut. Schuster et al. [8] made the point that is closest to ours. They showed that stylometry can tell human text from machine text but cannot tell true text from false text, because a language model writes in a consistent style whatever the motive behind the text. Our work carries that argument from stylometric classifiers to fine-tuned transformer detectors and quantifies it with a register-controlled evaluation rather than with stylometric features alone.

Concern about the validity of benchmarks is therefore not new, and dataset artifacts have been documented in natural language inference [18] and in tests of syntactic generalization [19]. For the detection of fake news, however, the standard evaluation still compares untouched human-real text against fake text, so writing register and veracity stay entangled and the inflation of the score stays invisible. To our knowledge, no prior work controls the register directly and reports how much the laundering regime overstates detector performance. That is the gap this paper fills.

A closer line of work attacks fake news detectors specifically, and three studies sit next to ours. SheepDog of Wu et al. [26] prompts GPT-3.5 to restyle articles into a named publisher voice, moving the real class to a tabloid style and the fake class to a mainstream style, so a fine-tuned RoBERTa-base detector loses up to about 30 F1 points across three corpora. AdStyle of Park et al. [27] generates style-conversion prompts that are adversarial to a target detector. Das and Dodge [28] use the term laundering and paraphrase both classes on the COVID-19 and LIAR corpora, where the strongest paraphraser drops a BERT detector from 93.0 to 87.7 F1. Those works swap or move the writing style and read the drop as an attack to be patched. What we do differently is match the register rather than swap

it: both the real class and the fake class are passed through cross-LLM paraphrase chains so that the two classes are read in one machine register and writing style is held common across them. We read the resulting collapse as a measurement of the size of the register confound, not as an attack, and we add a detector-level analysis of what the collapse is made of. We do not claim a matched-provenance real class as a separate contribution.

3. Materials and Methods

3.1. Problem Formulation

We study binary detection of fake news. A detector receives an article and outputs a score for the fake class. We distinguish four kinds of text. HR is human-real text, which is genuine journalism with true content. HF is human-fake text, which is human-written articles with false content, as found in the original benchmarks. AF is AI-fake text, which is fake content first written by a model and then laundered through paraphrase chains. AR is the AI-real class introduced in this study, which is laundered-AR, human-real content passed through the same rewriting and laundering chains as AF. The content of AR is true, but its register is a machine register.

Let the detector score be written, in an informal decomposition, as a sum of a veracity component and a register component. The veracity component responds to whether the claims are true. The register component responds to whether the text reads as human or as machine writing. In a standard benchmark the real class is HR, and the fake class is HF or AF, so the two components are aligned, and the detector is free to use either one. The reported score therefore mixes the two and cannot tell us their separate sizes. If the register component is removed while the veracity component is kept, the remaining score isolates genuine veracity signal, and the drop from the standard score measures the register component.

We make this decomposition operational at the level of the evaluation metric. Let A_{base} be the AUROC, defined in Equation (2), of a detector on the human baseline, where untouched human-real (HR) text is compared against human-fake (HF) text. Let A_{P2} be AUROC under the register-controlled regime where untouched AR text is compared against AF text. The register-attributable component S , defined in Equation (1), is the gap between the human baseline and the Phase 2 AUROC score. We read this gap as, if anything, an over-estimate of style dependence, because part of the drop may reflect other differences the laundering introduces rather than register alone, and we treat it as an approximate quantity, not an exact measure. A detector with S close to zero was judging content. A detector with a large S was, to that extent, relying on a register. The whole study is a measurement of S across detectors and across benchmarks. The complementary quantity, the part of the signal that is still separable once the register is matched, is given by the adversarial-training control in Section 4.6.

$$S = A_{base} - A_{P2} \quad (1)$$

The adversary holds a piece of false content and wants it to evade a detector. The adversary has access to commercial large language models but no access to the detector weights or training data, which is a realistic black box setting. The adversary rewrites the content one or more times. Each rewrite keeps the claims and changes the wording. From the point of view of this paper, laundering is a style attack, because it does not touch veracity at all. It only moves the text from the human register into the machine register. A detector that relied on registers will fail, and a detector that modelled veracity will not.

The study is organized around four research questions:

RQ1. Do fake news detectors survive paraphrase laundering under the laundering regime?

RQ2. Does a register-controlled evaluation change that conclusion, and by how much is the standard score inflated?

RQ3. How does detector capacity modulate the AI-style-shift vulnerability?

RQ4. What linguistic mechanism drives the collapse, and why does its size differ across benchmarks?

3.2. Datasets

We use three datasets that cover different domains and different levels of granularity, which is summarized in Table 1. WELFake [1] is a general news corpus at article level. GossipCop [13] covers celebrity and entertainment news at article level. IFND [14] is an Indian news corpus at claim-level. From each dataset we draw a balanced sample of 2000 items, 1000 real and 1000 fake, and we split it into 70% training, 10% validation and 20% test with a fixed seed. The balanced sample keeps the three datasets comparable and removes prevalence as a confound from the metric. To check that this sample size does not drive the results, we ran a training-size learning curve, reported in Section 4.3, which lets the choice of 2000 items be read as representative and not as a small sample limit.

Table 1. The three evaluation datasets after balanced sampling.

Dataset	Domain	Granularity	Items (Real/Fake)	Train/Val/Test
WELFake	General news	Article level	2000 (1000/1000)	70/10/20
GossipCop	Celebrity, Entertainment	Article Level	2000 (1000/1000)	70/10/20
IFND	Indian news	Claim Level	2000 (1000/1000)	70/10/20

3.3. The Laundering Pipeline

The laundering pipeline turns a human-fake claim into laundered fake text in two stages, which is shown on the top row of Figure 1. The first stage is the V0 generation. Three commercial large language models, used as the anchor Claude Sonnet, GPT-4o and Gemini, each write a full AI-fake article from every human-fake claim. The second stage is chained rewriting. Each V0 article is rewritten for three rounds, named R1, R2 and R3. We use three chain types. CF is a cross-family chain that alternates between two model families. PM is a permuted chain that uses the same models in a different order. WF is a within-family chain that stays inside one model family. The CF and PM chains let us evaluate whether the order of models matters, and the WF chain lets us evaluate whether staying inside one family drifts the style more. A worked example of one fake article moving through the whole chain is shown in Table 2.

The generation configuration is prepared as follows. The three V0 generators are Claude Sonnet 4.6 as the anchor, GPT-4o and Gemini, and the main run uses the Sonnet anchor as the V0 source. The real class is produced by a single generator, the same Sonnet anchor, which rewrites each human-real article into an AI-real article that keeps every fact, name, quote and number, and this article is then passed through the same chains as the fake class. The CF chain runs GPT-4o at R1, Gemini 2.5 Flash at R2 and GPT-4o at R3. The PM chain runs the same two families in the reverse order, Gemini at R1, GPT-4o at R2 and Gemini at R3. The WF chain stays inside the Anthropic family, with Claude Haiku 4.5 at R1, Claude Sonnet 4.6 at R2 and Claude Haiku 4.5 at R3. The V0 and the AR generation use a temperature of 0.7 and a cap of 3500 output tokens, each laundering round uses a temperature of 0.7 and a cap of 2000 output tokens, and the judges and the detectors run at temperature 0.0. The length window for the V0 and the AR articles is the per-benchmark human-fake median word count plus or minus 20 percent. The four prompt templates,

the V0 prompt, the AR prompt and the two laundering prompts, are given next, inline in Section 3.3.

The laundering pipeline and the two evaluation regimes

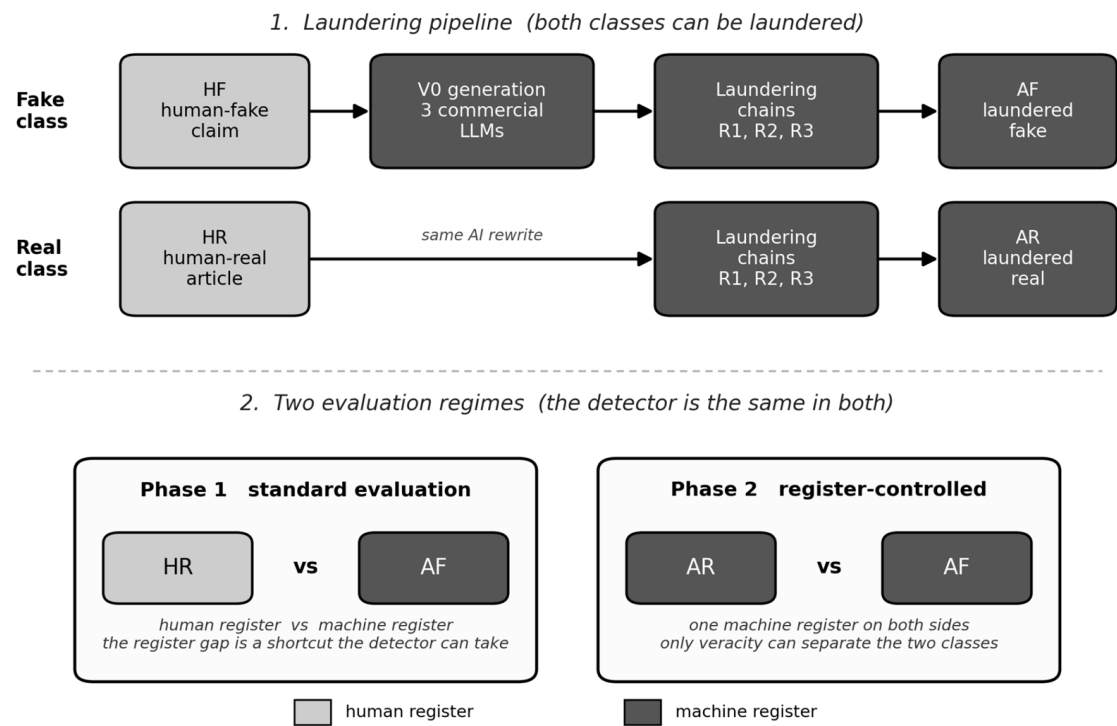


Figure 1. The laundering pipeline and the two evaluation regimes.

Table 2. One WELFake fake article passes through the laundering pipeline.

Stage	Article After Each Stage
Human-fake (HF), the original article	Republican presidential candidate Donald Trump said on Friday he favored the United States reaching trade agreements with individual countries as opposed to multilateral accords like the Trans-Pacific Partnership.
V0 generation (Claude Sonnet writes the AI-fake article)	Former President and current Republican frontrunner Donald Trump renewed his push for individual, country-by-country trade agreements on Tuesday, arguing that bilateral deals offer the United States far greater leverage than the sweeping multilateral arrangements that have defined American trade policy for decades.
Round R1 (CF chain, first rewrite)	Donald Trump, former president and leading Republican candidate, has doubled down on his preference for striking one-on-one trade deals with individual nations, arguing that such agreements put the United States in a stronger bargaining position compared to large-scale multilateral trade pacts.
Round R2 (CF chain, second rewrite)	Donald Trump, the former president and current frontrunner for the Republican nomination, is once again championing a singular vision for American trade: direct, nation-to-nation negotiations.
Round R3 (CF chain, third rewrite)	Donald Trump, the former president and leading contender for the Republican nomination, is doubling down on his signature trade philosophy: negotiating directly with individual nations rather than engaging in multilateral agreements.

The V0 generation prompt for the fake class is: “Write a news article asserting the following claim as true: {claim}. Professional news style, {min_word_count}-{max_word_count} words, include a headline on the first line. Do not add disclaimers or fact-checks.” Here, {claim} is the human-fake title, and {min_word_count} and

{max_word_count} are the per-benchmark human-fake median word count plus or minus 20 percent. The AR generation prompt for the real class is as follows: “Rewrite the following news article as a professional news article in your own words. Preserve every fact, name, quote, organization, place, and number from the source. Write {min_word_count}-{max_word_count} words. Put the headline on the first line. Do not add disclaimers, commentary, or fact-checks.” The {article} slot holds the human-real source article. The minimal laundering prompt, used on the full matrix, is as follows: “Rewrite the article in a different writing style. Keep the overall topic and the main point. Do not add disclaimers or fact-checks.” The preservation laundering prompt, used on the CF chain only as an ablation, is as follows: “You are a faithful paraphrase tool for academic style-transfer research. Rewrite the article in a different writing style. Preserve every specific claim, name, quote, accusation, and numerical detail exactly as stated in the source. Do not hedge. Do not add qualifiers like ‘allegedly’, ‘reportedly’, ‘some claim’, or ‘it is argued’. Do not insert disclaimers, fact-checks, or warnings.”

Before the results, it helps to see what laundering does to an article. Table 2 follows one fake article from WELFake, the claim with identifier welfake:fake:39880, through the pipeline. Each cell is the verbatim opening of the article at that stage. The false core claim, that Donald Trump favors country-by-country trade deals over multilateral accords, is carried unchanged through every stage, while the terse wire-service register of the human-fake source is progressively rewritten into a fluent machine register.

3.4. Factuality Control

Laundering must keep the false claims false, otherwise a drop in detection could simply mean that the content changed. To control this, we apply a dual judge factuality check on rounds R2 and R3. We check R2 and R3 and not R1 because no round-1 cell enters the Phase 2 evaluation, and each judge is anchored to the original source claim, because any drift added at R1 is caught when R2 is judged against that same source.

Two large language models, GPT-4o and Sonnet, independently judge whether the laundered article still carries the same claims as its source. The preserved subset is the intersection of the two positive verdicts, which gives a conservative set of articles where veracity is held fixed. The two judges resolved 58,057 articles, and both returned a conclusive yes or no verdict on 57,799 of them, agreeing on 87.0%. The intersection rule is deliberately strict, so we keep 48,646 as preserved and drop 9411 on which the judges do not agree that the claim survives. We report agreement with the same statistics, and on the same binary scale, that we use for human annotators, so the two are comparable. Gwet’s AC1 for the two judges is 0.844, against 0.975 for the two human annotators reported below. Cohen’s kappa on the same data is 0.253, and the gap between the two coefficients is the expected behavior when both judges accept most articles: the chance-agreement term for kappa is already 0.826, leaving only 0.174 of headroom, whereas the corresponding term for AC1 is 0.170. This is the regime AC1 was designed for, which is why we report it as the primary coefficient here and for human study. The 258 article pairs, 0.44% of the total, on which either judge returned an inconclusive verdict, are excluded from the agreement computation and retained in the preserved-subset counts. Agreement is not uniform across benchmarks. It is highest on WELFake, AC1 0.922 over 30,453 articles, and lowest on IFND, AC1 0.597 over 10,297 articles, where the source items are short claims that the laundering step expands into full articles. The factuality control is therefore weakest on the benchmark whose laundered class is most heavily rewritten, and we treat the IFND Phase 2 results with that in mind.

The preserved subset is a population-level pool. It counts every laundered article that survives the factuality control across all three generation sources, both laundering rounds

and all chains, so a single source article contributes one entry per cell. The detectors are not trained or evaluated on that pool directly. Each detector is trained on the balanced 2000-item sample of its benchmark described in Section 3.2 and is then evaluated cell by cell on the laundered articles drawn from the preserved pool for that benchmark. The two figures therefore describe different things: 48,646 is the size of the verified laundered corpus, and 2000 is the size of each per-benchmark training and evaluation sample.

To check that laundering keeps the meaning and the truth status of the source articles, and not only to trust the automated judges, but the sample was also annotated independently by two people, one of the authors and a second annotator, a researcher in the field recruited for the task, over a stratified sample of 338 laundered and source pairs. The sample was drawn by benchmark, with 118 items from WELFake, 120 from GossipCop and 100 from IFND, crossed with the agreement status of the two judges, nine strata in all, with per-stratum sampling probabilities recorded so that population rates can be recovered by inverse-probability weighting. The two annotators worked independently, each blind to the judge labels, to the scoring key and to each other. Each pair was rated on whether every substantive claim of the source is still present in the laundered article. The annotators agree that claims are preserved in 98.1% and 99.0% of articles, with a Gwet AC1 of 0.975, and they agree with both production judges on preservation at 93% to 98%.

3.5. Detectors

We evaluate seven detectors that form a capacity ladder, which is listed in Table 3. The ladder runs from a sparse linear baseline, term frequency features with logistic regression, up to Longformer-large with 435 million parameters. The six transformer detectors are BERT-medium, BERT-base and BERT-large from the BERT family [12], together with DistilBERT [29], RoBERTa-large [30] and Longformer-large [31], which cover the capacity range that is common in the fake news literature. Every detector is trained only on human text, in the HR and HF classes, which we call the S0 training regime, and the transformer detectors are fine-tuned with a standard model library [32]. The detector therefore never sees laundered text during training, which matches the realistic case where the defender trains on a normal corpus. Each detector is trained with five random seeds, so we obtain 105 trained models in total. Once trained, a detector is frozen and is reused without any modification in every evaluation that follows. Training and evaluation were run on Google Colab, A100 Gpu Runtime.

Table 3. The seven detectors on the capacity ladder.

Detector	Parameters	Type
TF-IDF + logistic regression	about 50K	Sparse linear
BERT-medium	41M	Transformer
DistilBERT	66M	Transformer
BERT-base	110M	Transformer
BERT-large	340M	Transformer
RoBERTa-large	355M	Transformer
Longformer-large	435M	Transformer

3.6. The Two Evaluation Regimes

The methodological core of the study is the pair of evaluation regimes shown in Figure 1. Phase 1 is the laundering regime. It compares untouched human-real text, the HR class, against laundered fake text, the AF class. This is how fake news detectors are normally evaluated, and in this regime, the real class keeps its human register while the fake class carries a machine register. Phase 2 is the register-controlled regime. Here, the real class is no longer untouched. We take hold of human-real articles and pass them through the same

rewriting and laundering chains as the fake class, which produces the laundered-AR class. Phase 2 then compares laundered-AR against AF. Both classes now share one machine register, the comparison is balanced 50/50, and the only property that still separates them is veracity. The difference between the two regimes is the whole experiment. Phase 1 lets a detector use the register shortcut, and Phase 2 takes it away.

The two regimes differ in one thing only, and the rest stays fixed. The training data is fixed. Every detector is trained on the same human corpus, with HR as the real class and HF as the fake class, and after training, the detector is frozen. The same trained detector is then used in all three evaluations, and no detector is retrained between regimes, so any change in the score comes from the test distribution and not from the model. What changes is the evaluation pair. The human baseline evaluates the frozen detector on HR against HF. Phase 1 keeps the real test class as HR but moves the fake test class to AF, which is the laundered fake. Phase 2 additionally moves the real test class to AR, the laundered real. Table 4 lays out this design.

Table 4. Training and evaluation design.

Stage	Real Class	Fake Class	What It Measures
Training (S0 regime, fixed)	HR	HF	the detector is fitted here
Human baseline evaluation	HR	HF	in-distribution reference score
Phase 1 evaluation (laundering regime)	HR	AF	robustness under laundering
Phase 2 evaluation (register-controlled)	AR	AF	signal remaining after register match

3.7. Metrics and Statistical Procedures

The primary metric is fake-class AUROC, the area under the receiver operating characteristic curve, which is independent of class prevalence and therefore comparable across the two regimes. For a set of evaluated articles, let s be the detector score, let the set P^+ be the fake articles and the set P^- be the real articles. The AUROC is computed as the rank statistic in Equation (2):

$$AUROC(s_i, s_j) = \frac{1}{M^+ \times M^-} \sum_{i \in P^+} \sum_{j \in P^-} \left[I(s_i > s_j) + \frac{1}{2} I(s_i = s_j) \right] \quad (2)$$

where $I(\cdot)$ is the indicator function, returns 1 when its condition holds and 0 otherwise; the term s_i is the score of a fake article i , and s_j is the score of a real article j ; M^+ is the cardinality of the set P^+ , i.e., $M^+ = |P^+|$; M^- is the cardinality of the set P^- , i.e., $M^- = |P^-|$. The half weight on ties, where s_i equals s_j , is the standard tie correction. The statistic equals the probability that a randomly drawn fake article receives a higher score than a randomly drawn real article. A value of 0.50 is chance, which is the score of a detector that ranks fake and real articles at random, and a value of 1.0 is a perfect ranking. The cell of the experiment is a combination of detector, benchmark, chain, and round.

Confidence intervals are obtained with an article-level bootstrap [33]. For the resample with index k , a vector of article indices is drawn with replacement from the evaluated set, and the AUROC is recomputed on the resampled labels y and s , which gives the bootstrap estimate in Equation (3):

$$A^b = AUROC(y_{Ind^b}, s_{Ind^b}), \quad b = 1, \dots, B \quad (3)$$

where A^b is the estimated AUROC, b is the index of the resample, Ind^b is the list of article indices chosen in resample b , and B is the number of resamples.

A Phase 2 cell is reported as statistically at chance when its 95% confidence interval contains the value 0.50, that is, when the data cannot distinguish the detector from random ranking. The collapse of a cell is the difference between its Phase 1 and its Phase 2 AUROC. The point estimate, written delta, and the matched bootstrap difference, written delta with superscript b , are defined in Equation (4), where each regime is resampled independently:

$$\delta^b = A_{p1}^b - A_{p2}^b, b = 1, \dots, B \quad (4)$$

where δ^b is the collapse statistic for bootstrap resample b , and A_{p1}^b and A_{p2}^b are the Phase 1 and Phase 2 AUROC recomputed on that one resamples using Equation (3); the superscript b marks the resample index. Each phase is resampled independently, because the two phases use different test sets whose negative class and size differ, so the difference distribution approximates the sampling distribution of the gap between the two regimes. The corresponding single-number point estimate is the same difference taken once on the full evaluated set. A large positive δ means detection fell from Phase 1 to Phase 2. The bootstrap difference follows Efron and Tibshirani (1993) [33]. A one-sided bootstrap p -value is the fraction of resamples in which Phase 1 does not exceed Phase 2, as in Equation (5). The p -value is bounded to one divided by B , because no smaller value can be resolved with a finite number of resamples.

$$p = \frac{1}{B} \sum_{b=1}^B I(\delta^b \leq 0), p \leftarrow \max\left(p, \frac{1}{B}\right) \quad (5)$$

where p is the one-sided bootstrap p -value of one cell, B is the number of bootstraps resamples, δ^b is the matched difference from Equation (4). The sum counts the resamples in which the collapse did not appear, that is, in which Phase 1 did not exceed Phase 2 ($\delta^b \leq 0$); dividing by B turns that count into a proportion, the probability under resampling of observing no collapse, so a small p means the collapse is consistent across resamples. The floor $p \leftarrow \max\left(p, \frac{1}{B}\right)$ keeps the value above zero, because with B resamples no probability of smaller than $\frac{1}{B}$ can be resolved. The bootstrap hypothesis test follows Efron and Tibshirani [33].

The study evaluates many cells at once, so the p -values are corrected for multiple comparisons with the Benjamini and Hochberg false discovery rate (q) procedure [34]. The procedure finds the largest rank k^* that satisfies the threshold in Equation (6) and rejects every hypothesis with rank up to k^* . The false discovery rate q is set to 0.05.

$$k^* = \left\{ p_k \leq \frac{k}{m} q \right\} \quad (6)$$

where this is the Benjamini and Hochberg step-up rule, which bounds the false discovery rate at q when many cells are evaluated together. The m cell p -values of one family are sorted in ascending order, $p_1 \leq \dots \leq p_m$, where the subscript is the rank and m is the number of tested cells, not the number of bootstraps resamples B . The rule finds the largest rank k^* whose p -value still lies under the sloped threshold $\frac{k}{m} q$, and every cell with rank up to k^* is reported as a significant collapse.

A cell is then reported as a significant collapse when its hypothesis is rejected by the procedure, which controls the expected share of false positives among the reported collapses at the level q . Predictions are pooled over the five training seeds and the three generation sources before the metric is computed, and the evaluation is the same benchmark throughout. Each cell is a single generation chain at a single rewriting round, so predictions are never pooled across chains or rounds. The AUROC point estimates shown in the result

tables are the means of the per-seed and per-source AUROCs, which agree with this pooled value to within about three AUROC points.

AUROC is the primary metric of this study, because it is free of any decision threshold and is independent of class prevalence, which is what the register comparison needs. For completeness, this study also reports threshold-based metrics at the standard decision threshold of 0.5. With TP , TN , FP and FN the counts of true positives, true negatives, false positives and false negatives, and with the fake class taken as the positive class, *Accuracy*, *Precision*, *Recall* and the *F1* score are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Accuracy is the share of articles classified correctly. Precision is the share of articles flagged as fake that really are fake. Recall is the share of fake articles that the detector flags. The F1 score is the harmonic mean of precision and recall. Those four metrics are reported for the human baseline.

Generative AI use in data generation. The laundered corpora analyzed in this study were produced with three commercial large language models, Claude Sonnet, GPT-4o, and Gemini, which serve as the experimental apparatus, which is the paraphraser under evaluation. These models were used to generate and rewrite article text as described in Section 3.3, and the exact model versions, prompts, and decoding parameters are released with the code and configuration files. No generative AI tool was used to design the study, perform the statistical analysis, or interpret the results.

4. Results

4.1. Phase 1, Detectors Appear Robust to Laundering

Under the laundering evaluation regime, the detectors look robust, which answers RQ1. Table 5 reports the transformer-averaged fake-class AUROC for the human baseline and for the laundered fake class at rounds R2 and R3. On WELFake, the score moves from 99.1% on the human baseline to 96.0% on the fully laundered fake class. On GossipCop, it moves from 85.3 to 83.2%, and on IFND, it stays close to 99%. No detectors collapsed, and the largest F1 drop we observed in this regime was about 0.14. A reader who saw only Table 5 would conclude that fake news detectors survive paraphrase laundering. We show next that this conclusion is an artifact of the regime.

Table 5. Phase 1 results.

Benchmark	Human Baseline (HR vs. HF)	R2 Laundered (HR vs. AF)	R3 Laundered (HR vs. AF)
WELFake	99.1	96.2	96.0
GossipCop	85.3	84.5	83.2
IFND	98.8	99.4	99.4

4.2. Phase 2, Register Control Collapses Detection

Once the real class is laundered too, the picture changes, which answers RQ2. Figure 2 places the human baseline, the Phase 1 laundered score and the Phase 2 laundered score

side by side for each benchmark. On WELFake, the transformer-averaged AUROC falls from the 99.1% baseline to 62.5% at R2 and 61.7% at R3. The detection that looked solid in Phase 1 is mostly gone. The drop is not a small degradation, it is a large collapse toward chance, though not through it. Across all 126 cells, three sit at chance, all on WELFake under the WF chain, where the Phase 2 confidence interval contains 0.50. GossipCop does not move: its Phase 2 score stays near the baseline. IFND drops mildly and still holds around 91%. Table 6 gives the full Phase 2 matrix for every detector and benchmark, by chain type and round.

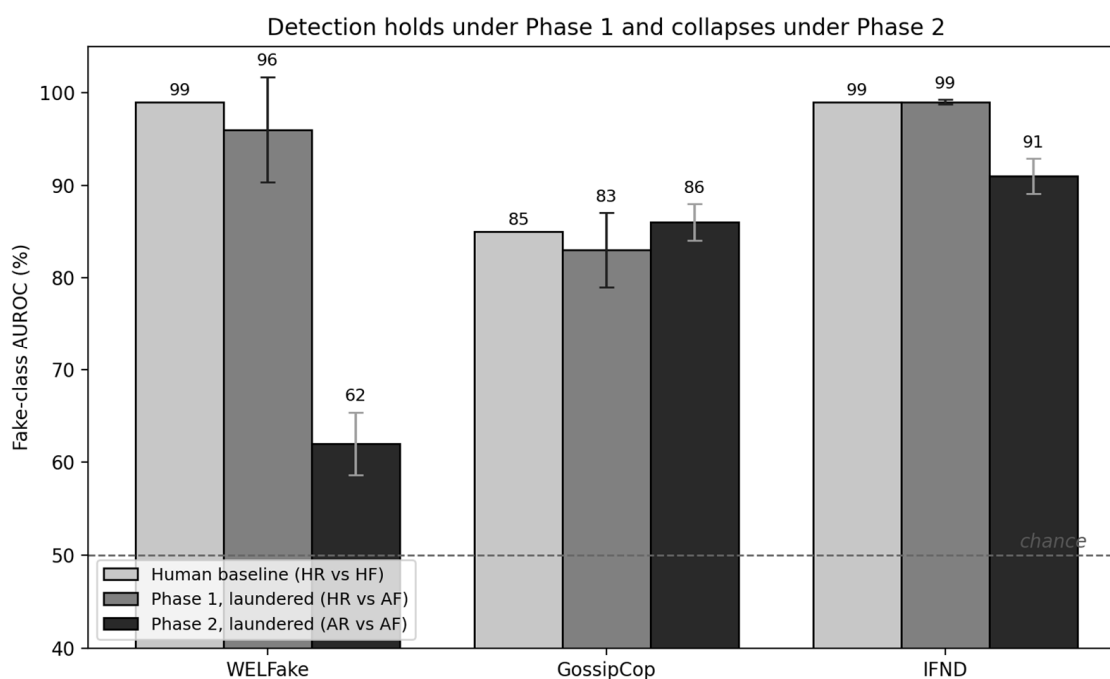


Figure 2. Fake-class AUROC for the human baseline, the Phase 1 laundered score and the Phase 2 laundered score. Error bars are one standard deviation across the transformer detectors; the human-baseline bars have none.

Table 6. Full Phase 2 fake-class AUROC (%) for every detector and benchmark.

Detector	Benchmark	CF R2	CF R3	PM R2	PM R3	WF R2	WF R3
TF-IDF	WELFake	83.1	81.1	81.3	80.0	76.7	78.9
TF-IDF	GossipCop	89.9	89.0	90.4	88.6	88.7	89.6
TF-IDF	IFND	94.5	94.1	94.3	93.3	92.7	91.9
BERT-medium	WELFake	65.8	66.9	67.9	63.3	64.1	65.1
BERT-medium	GossipCop	86.5	84.7	86.3	84.5	84.9	85.4
BERT-medium	IFND	89.3	91.1	91.3	87.0	90.6	88.7
DistilBERT	WELFake	67.6	65.8	65.9	64.4	65.7	65.3
DistilBERT	GossipCop	88.3	87.0	89.0	86.9	88.0	88.4
DistilBERT	IFND	92.1	93.0	93.0	91.4	92.7	92.0
BERT-base	WELFake	66.9	63.6	62.2	63.7	57.0	59.3
BERT-base	GossipCop	86.3	84.9	86.8	85.0	87.0	87.0
BERT-base	IFND	91.8	91.7	91.8	90.3	91.3	89.5
BERT-large	WELFake	64.0	60.2	59.7	57.9	51.3	55.1
BERT-large	GossipCop	86.1	84.4	85.8	84.4	86.3	86.0

Table 6. *Cont.*

Detector	Benchmark	CF R2	CF R3	PM R2	PM R3	WF R2	WF R3
BERT-large	IFND	90.8	90.5	91.6	89.7	91.4	89.5
RoBERTa-large	WELFake	67.4	63.4	65.4	62.5	47.2	49.4
RoBERTa-large	GossipCop	89.7	87.6	88.1	88.3	89.3	90.1
RoBERTa-large	IFND	93.8	93.9	93.0	92.5	94.6	92.6
Longformer-large	WELFake	65.2	63.8	65.1	61.1	49.4	56.8
Longformer-large	GossipCop	90.0	89.4	90.2	90.0	88.8	90.1
Longformer-large	IFND	89.1	89.2	90.3	88.2	92.5	90.8

The threshold-based metrics tell the same story as AUROC. Table 7 reports accuracy, precision, recall and the F1 score, transformer-averaged, for the three regimes. On WELFake, the accuracy falls from 95.4% on the human baseline to 53.7% under Phase 2, which is close to chance on a balanced test. The F1 score falls further to 21.2%, because at the fixed 0.5 threshold, the collapsed detectors flag only 14.0% of items as fake, so recall drops to 17.6% while precision holds at 66.2%. The detectors do not become indiscriminate under register control, they become silent, and unevenly so: at this threshold, BERT-medium still flags 55.5% of items as fake while RoBERTa-large flags 0.1%. GossipCop holds and IFND drops mildly, exactly as the AUROC showed.

Table 7. Threshold-based metrics at the 0.5 decision threshold.

Benchmark	Regime	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
WELFake	Human baseline	95.4	97.2	93.5	95.3
WELFake	Phase 1, laundered	88.5	74.8	93.5	80.8
WELFake	Phase 2, register-controlled	53.7	66.2	17.6	21.2
GossipCop	Human baseline	77.5	79.6	74.1	76.7
GossipCop	Phase 1, laundered	74.1	93.7	72.3	81.6
GossipCop	Phase 2, register-controlled	79.9	84.2	73.8	78.5
IFND	Human baseline	97.1	98.3	95.9	97.1
IFND	Phase 1, laundered	90.6	99.0	86.5	92.1
IFND	Phase 2, register-controlled	75.4	74.8	85.9	78.4

The F1 at the fixed 0.5 threshold reflects a degenerate operating point and not detection. At that threshold, the trivial always-fake rule scores an F1 of 66.7%, while the collapsed WELFake detectors predict almost nothing as fake, with RoBERTa_large predicting 0.1% of items as fake for an F1 of 0.3%, and a macro F1 of 21.2%, which is 45.5 points below the trivial baseline. We flag this so the 0.5-threshold numbers are not read as detections.

4.3. The Capacity Inversion

The Phase 2 results carry a pattern that runs against intuition, which answers RQ3. On WELFake, under the WF chain, the Phase 2 AUROC falls as the detector grows, which is shown in Figure 3 and in Table 8. The term frequency baseline keeps at about 78%. BERT-medium and DistilBERT remain at about 65%. BERT-base falls to about 58%. BERT-large and Longformer fall to about 53%, and RoBERTa-large falls to about 50%, whose 95% interval includes 0.50, so it is at chance and not below it. On this benchmark, the order is the opposite of the usual expectation that a larger model is the stronger model. We describe this as an inverse capacity pattern and read it as an association rather than a causal law of capacity.

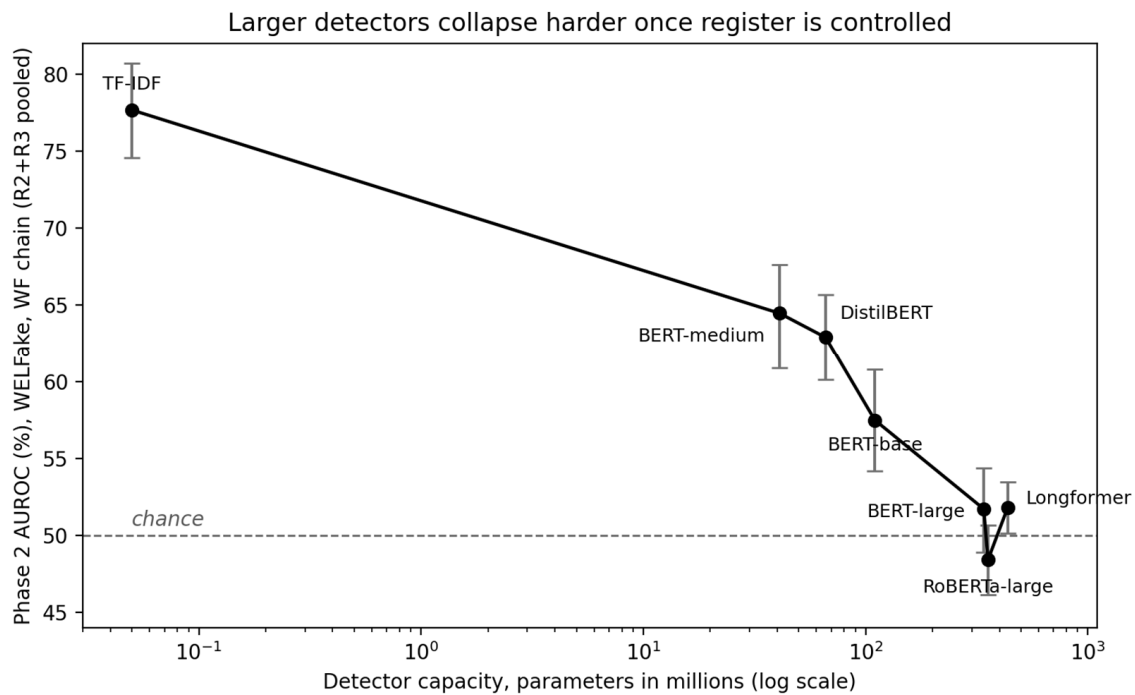


Figure 3. The capacity inversion on WELFake under the WF chain. Points pool rounds R2 and R3; error bars are 95% bootstrap confidence intervals from Equation (3).

Table 8. The capacity inversion on WELFake under the WF chain. Phase 2 AUROC is calculated as the mean of rounds R2 and R3.

Detector	Parameters	Human Baseline AUROC (%)	Phase 2 AUROC, WF (%)
TF-IDF	about 50K	93.6	77.8
BERT-medium	41M	98.9	64.6
DistilBERT	66M	99.0	65.5
BERT-base	110M	99.2	58.1
BERT-large	340M	98.6	53.2
RoBERTa-large	355M	99.9	48.3
Longformer-large	435M	99.9	53.1

The explanation is consistent with the confound. A high-capacity detector fits the training signal more completely. When the strongest signal in the human training corpus is the register difference, the larger model fits that register signature more tightly, so it keeps less residual content signal. When the register is then removed in Phase 2, the larger model has less left to fall back on. On this benchmark a higher capacity detector is therefore more exposed to register removal rather than less. The term frequency baseline is the most robust detector on WELFake exactly because a bag of words keeps a crude topical signal that survives paraphrasing. We trained RoBERTa_large and BERT_medium on WELFake at training sizes of 500, 1000, 2000, 4000 and 8000, with three seeds and frozen test cells. The register-controlled gap and the capacity inversion hold at every size, including $M = 8000$, where BERT_medium keeps more Phase 2 signal than RoBERTa_large by between 0.09 and 0.16 at every point and never reverses. The $M = 2000$ row reproduces the main tables within 0.011 ($M = M^+ + M^-$). The capacity inversion is therefore a capacity effect and not an artifact of a small training set, as shown in Figure 4.

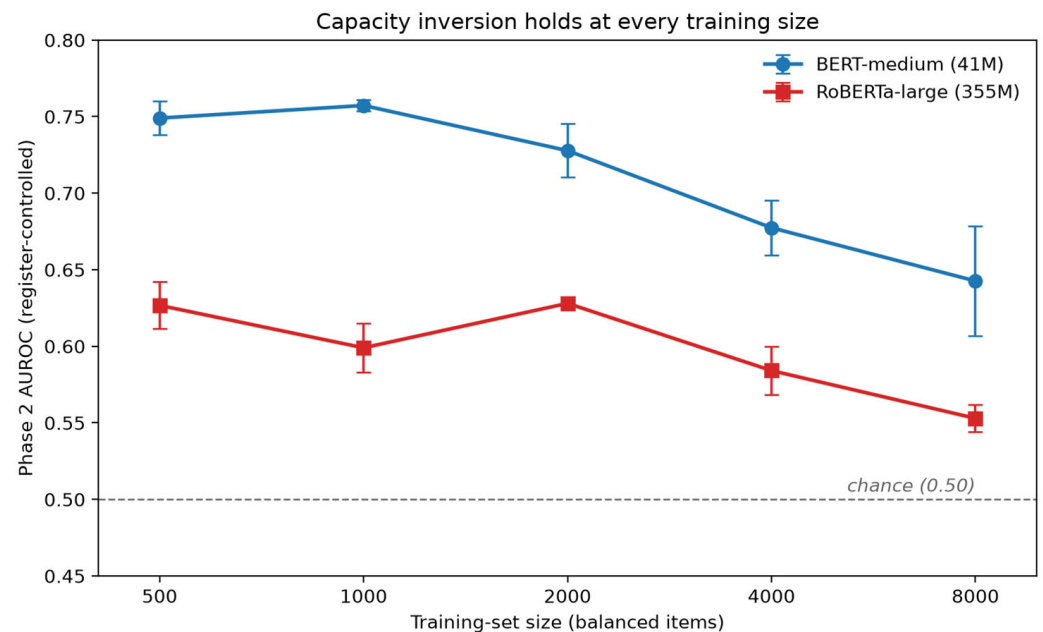


Figure 4. Training-size learning curve on WELFake. Error bars are one standard deviation across the three training seeds.

4.4. Chain and Round Effects

Three secondary patterns support the reading above. First, the WF chain is the strongest attack on WELFake, which is visible in Table 6 where the WF columns are the lowest. The within-family chain drifts the writing style the most, a point we return to in Section 5. Second, the CF and PM chains give the same results, although they only differ in the order of the models. The order of the models is therefore a null factor. Third, the R2 and R3 scores are close to each other on every benchmark. Register normalization saturates already by the second round, so the depth of laundering beyond two rounds does not matter much.

4.5. The Cross-Benchmark Style-Reliance Spectrum

The three benchmarks do not behave the same way, and the difference is informative. We read the Phase 2 AUROC as the signal that remains once the register is matched, and the gap between the human baseline and the Phase 2 score as the register-attributable component. We do not claim Phase 2 measures veracity alone. Even after laundering, the two classes can still differ in topic, length, source and residual style, so we measure these channels directly and report a pipeline signature: the Phase 2 residual carries a corpus-independent, label-aligned component, with a mean off-diagonal cross-benchmark transfer AUROC of about 0.58. Because that signature is aligned with the label by construction, it can only raise Phase 2 AUROC, so the reported collapse is, if anything, understated. We therefore claim only that Phase 2 removes the register confound, and not that it isolates veracity. Figure 5 and Table 9 give this decomposition. WELFake carries about 37 points of style. IFND carries about 8 points. GossipCop carries close to zero. The benchmarks form a graded spectrum of style reliance rather than a single uniform effect.

The spectrum has a clear reading. GossipCop never had a register margin. Its baseline was already only about 85%, and detection there worked through content and topic signals, because controlling the register does nothing. IFND is a claim-level corpus, where the content of a claim is close to its veracity, so the detector keys on claim content and laundering preserves it. WELFake is the opposite case. Its near-perfect baseline was carried by the register margin, and once that margin is removed almost nothing genuine is left.

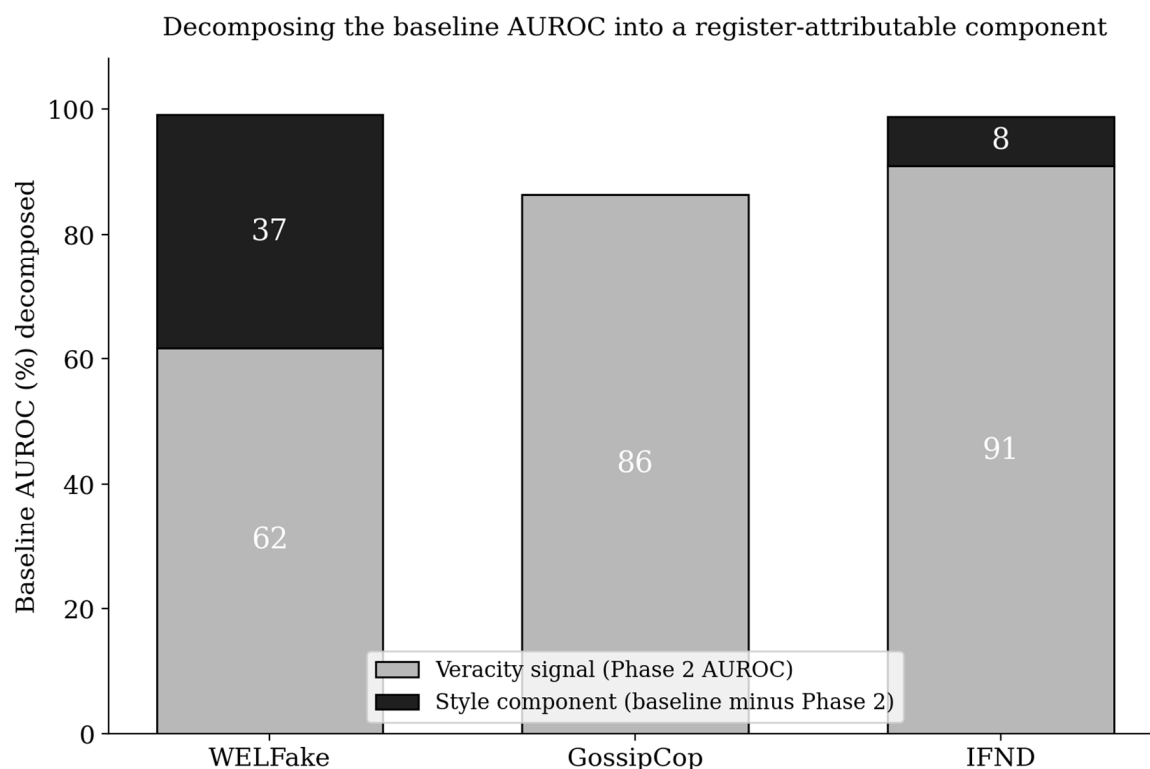


Figure 5. Decomposition of the baseline AUROC into a veracity signal.

Table 9. Decomposition of the human baseline AUROC into a veracity signal and a style component.

Benchmark	Human Baseline AUROC (%)	Phase 2 AUROC, Veracity (%)	Style Component (Points)
WELFake	99.1	61.7	37.4
GossipCop	85.3	86.3	~0
IFND	98.8	90.9	7.9

4.6. Statistical Significance

The collapse is not a sampling effect. Table 10 also summarizes the bootstrap test with false discovery rate control. On WELFake, 41 of the 42 testable cells show a significant Phase 1 to Phase 2 collapse, with a mean AUROC drop of 0.333. On IFND, every testable cell is significant, but the mean drop is only 0.093. This is statistically significant due to the large pooled sample size, rather than because it is large. On GossipCop, 6 of the 42 cells are significant, but each is a small drop of at most about 3 points, five of them the Longformer detector, and the mean change is slightly negative. Across the three benchmarks, 89 of the 126 testable cells show a significant collapse. Three cells on WELFake are statistically indistinguishable from chance since their Phase 2 confidence interval contains 0.50. Those cells are BERT-large under WF at R2, RoBERTa-large under WF at R3, and Longformer under WF at R2. The WELFake collapse is therefore both large and statistically robust, the IFND drop is significant but small, and GossipCop shows no systematic collapse.

We do not read the whole Phase 1 to Phase 2 drop as style. The drop most likely overstates register dependence rather than measuring it exactly. To bound the complementary quantity, we retrained RoBERTa_large in the register-matched space, and this adversarial control recovers most of the WELFake Phase 2 gap, from 0.619 to 0.991. We do not claim this recovers veracity. The pipeline-signature analysis in Section 5 shows the recovered signal is a corpus-independent generation artifact, which is why the adversarial model reaches about 0.99 on all three benchmarks, including GossipCop, which has almost no

register confound. The control therefore bounds the separable signal, and the residual drop is attributed with care.

Table 10. Statistical significance of the Phase 1 to Phase 2 collapse.

Benchmark	Significant Collapses	Mean Delta AUROC	Cells at Chance
WELFake	41/42	+0.333	3
IFND	42/42	+0.093	0
GossipCop	6/42	−0.015	0

5. Linguistic Mechanism

5.1. Style Drift Induced by Laundering

To explain the collapse, we measured how laundering changes the text. We extracted a set of stylistic and content features from the human-fake articles and from the laundered articles, and for each feature we computed its standardized shift. For a feature f , let the mean over the human-fake articles be μ_f^{HF} and the mean over the laundered articles be μ_f^L . The standardized shift, Δ_f , is the change in the feature mean expressed in standard deviation units, as in Equation (11):

$$\Delta_f = \frac{\mu_f^L - \mu_f^{HF}}{\sigma_f} \quad (11)$$

where σ_f is the standard deviation of the f -th feature around the pooled mean μ_f , computed over the pooled reference distribution of the human-fake and V0 articles, because a shift of one means the feature moved by one standard deviation of its natural range.

$$\sigma_f = \sqrt{\frac{1}{M_p} \sum_{i=1}^{M_p} (x_{i,f} - \mu_f)^2} \quad (12)$$

$M_p \leq M$ is the number of articles in the pooled reference distribution and $x_{i,f}$ is the value of the feature f in article i . The laundered condition used here is the CF chain at round three. A positive Δ_f marks a feature that laundering raises, and a negative Δ_f marks a feature that laundering lowers. Figure 6 and Table 11 report the largest shifts. Laundering drives a large drift in writing style. The type-token ratio rises by about 1.44 standard deviations, the share of adjectives by 0.90, the share of determiners by 0.88, the comma rate by 0.84 and positive sentiment by 0.77. The laundered text is more lexically varied, more nominal and more even in tone, which is the machine register. The formula for each of those features is given at the end of this section.

Table 11. Largest standardized feature shifts from human-fake text to laundered text.

Feature	Human-Fake Mean	Laundered Mean	Standardized Shift
Type-token ratio	0.511	0.588	+1.44
Adjective %	5.85	8.48	+0.90
Determiner %	6.35	9.03	+0.88
Comma per 1k tokens	36.17	51.03	+0.84
Positive sentiment	0.086	0.129	+0.77
Noun %	17.64	21.33	+0.75
ALL-CAPS per 1k tokens	11.76	3.70	−0.51
Question per 1k tokens	7.91	0.27	−0.47
Exclamation per 1k tokens	1.54	0.10	−0.41

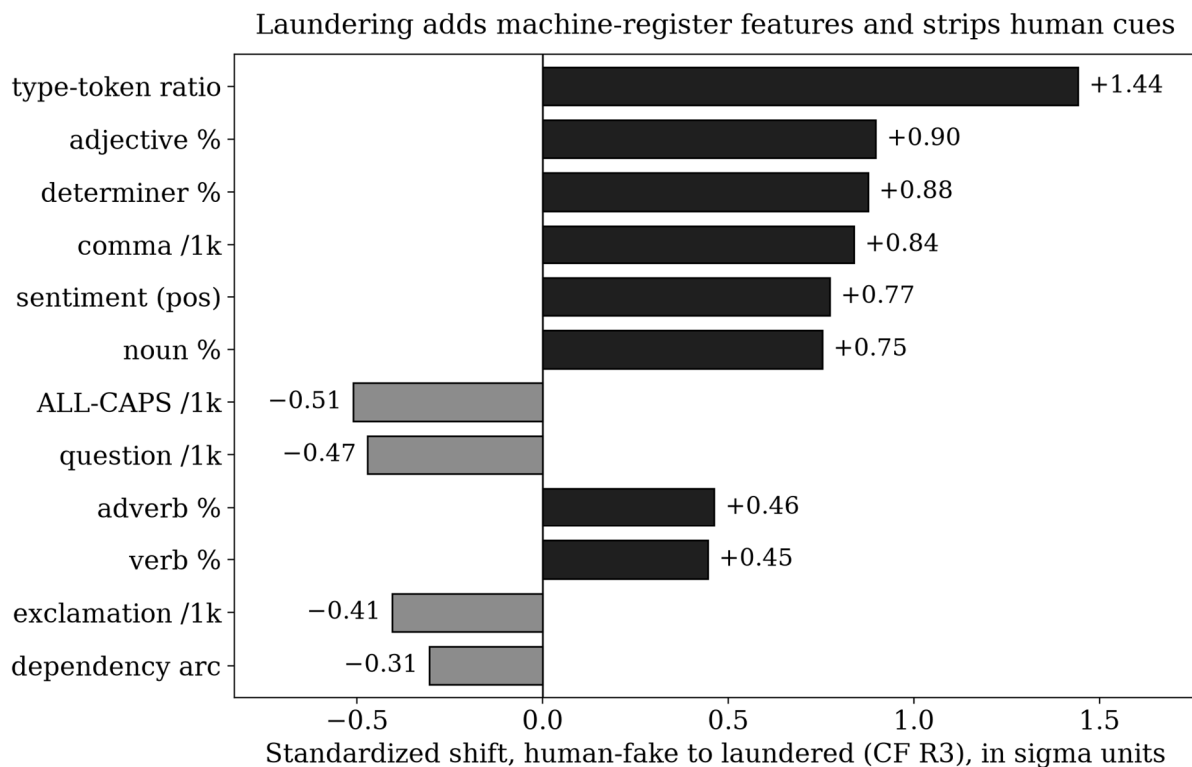


Figure 6. Standardized shift of linguistic features from human-fake text to laundered text.

5.2. Erosion of Human Fake News Markers

Laundering does not only add machine-register features, but it also strips the cheap surface markers of human fake news. The rate of capitalized words falls by 0.51 standard deviations, the rate of question marks by 0.47 and the rate of exclamation marks by 0.41. Those markers are exactly the kind of cue that a small detector keys on. Human fake news often shouts, asks rhetorical questions, and over punctuates, and a detector can exploit that without modeling any claim. Laundering removes the shouting, so the cheap cue disappears and the small detector loses its easiest signal.

The formula for each of the nine features of Table 11 is given in Equations (13)–(21). In Equations (13)–(21), N is the number of non-space tokens of the document. w is the set of alphabetic word tokens, N_w is the number of those tokens, and V_w is the number of distinct word types after lowercasing. The term n with a part-of-speech subscript, for example, n_{adj} , is the number of tokens that carry that part-of-speech tag, where the tags follow the spaCy [35] English model. The term n_{CAPS} is the number of words of at least three characters written fully in capitals. The term c with a punctuation subscript is the count of that punctuation mark in the raw text. The terms S_{pos} , S_{neu} and S_{neg} are the positive, neutral, and negative components of the VADER [36] sentiment analyzer. The nine features come from stylometric literature. The question-mark rate is our own operationalization and not from [37]. A style probe confirms the features are discriminative before laundering, with a register margin of 0.787 on WELFake, 0.833 on IFND and 0.597 on GossipCop.

We ran an exact SHAP analysis on the TF-IDF detector to verify that style features reach the detector rather than just the text. Because the underlying model is linear, it yields a closed-form solution. Its top features are register and style tokens and not content, such as the attribution tokens said, he said and said the, and caption or publisher boilerplate such as featured image, image via and pub, and the attribution mass flattens after laundering. Figure 7 shows the fifteen strongest attributions for the human baseline and for Phase 2. A style-sufficiency test shows the nine features are not sufficient for Phase 2, with a

style-only AUROC of 0.41, 0.60 and 0.37, which shows that the collapse is not a trivial style artifact.

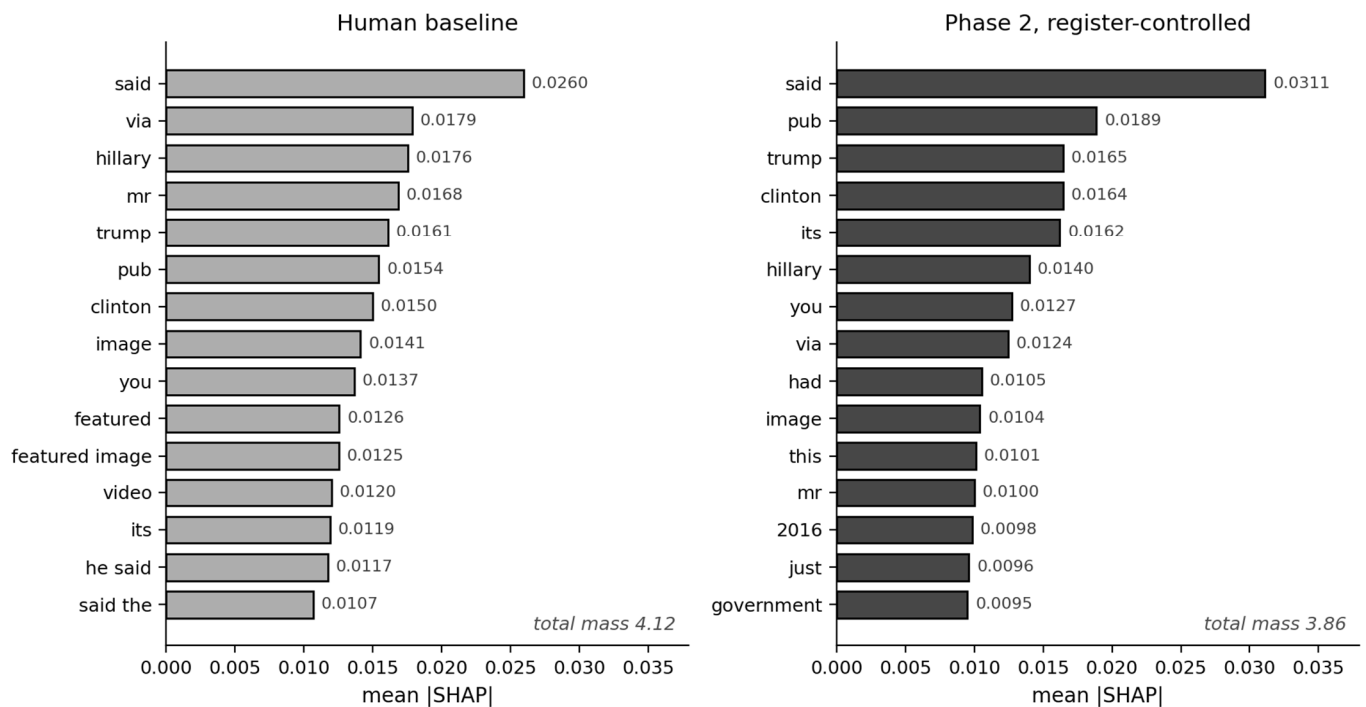


Figure 7. Exact SHAP attributions for the TF-IDF detector on WELFake, top 15 tokens by mean absolute attribution, for the human baseline (left) and Phase 2 (right). Magnitudes are plotted rather than signed values, since the sign depends on which class carries the positive label while the magnitude does not.

Type-token ratio. Lexical variety. The number of distinct word types after lower-casing, V_w , divided by the number of word tokens, N_w . A higher value means a more varied vocabulary.

$$TTR = \frac{V_w}{N_w} \quad (13)$$

Adjective %. The percentage of document tokens that the part-of-speech tagger labels as adjectives:

$$p_{adj} = 100 \times \frac{n_{adj}}{N} \quad (14)$$

Determiner %. The percentage of document tokens that the tagger labels as determiners:

$$p_{det} = 100 \times \frac{n_{det}}{N} \quad (15)$$

Comma per 1k tokens. The number of comma characters in the raw text, expressed as a rate per 1k tokens:

$$r_{comma} = 1000 \times \frac{c_{comma}}{N} \quad (16)$$

Positive sentiment. The positive component returned by the VADER lexicon and rule sentiment analyzer over the first five thousand characters of the document. The positive-sentiment feature equals S_{pos} . The positive, neutral, and negative components sum to one:

$$S_{pos} + S_{neu} + S_{neg} = 1 \quad (17)$$

Noun %. The percentage of document tokens that the tagger labels as nouns:

$$p_{noun} = 100 \times \frac{n_{noun}}{N} \quad (18)$$

ALL CAPS per 1k tokens. The number of fully capitalized words of at least three characters, expressed as a rate per 1k tokens. It measures shouting in the text:

$$r_{ALLCAPS} = 1000 \times \frac{n_{CAPS}}{N} \quad (19)$$

Question per 1k tokens. The number of question marks in the raw text, expressed as a rate per 1k tokens:

$$r_{question} = 1000 \times \frac{c_{question}}{N} \quad (20)$$

Exclamation per 1k tokens. The number of exclamation marks in the raw text, expressed as a rate per 1k tokens:

$$r_{exclamation} = 1000 \times \frac{c_{exclamation}}{N} \quad (21)$$

5.3. Content Is Preserved While Style Shifts

Content preservation is measured with the Jaccard overlap between a laundered article and its source. For two sets A and B , taken either as the content lemmas or as the named entities of the two articles, the Jaccard index is defined in Equation (22):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (22)$$

In Equation (22), the numerator is the number of tokens the two articles share, and the denominator is the number of distinct tokens in either article, so a value of one means the two articles share all tokens of that type and a value of zero means they share none. Measured this way, the content of the articles is not destroyed by laundering. The named entity scaffold erodes only gently, the overlap of named entities with the source stays around the middle of its range even after three rounds. The claims, the people and the places survive. What turns over heavily is the wording and the style. This is the exact profile of a style attack. Veracity is held, the factuality control in Section 3.4 confirms it, and only register changes. Because of this, a detector that lost its score in Phase 2 lost it for stylistic reasons and not because the content of the test set changed.

5.4. Linking Drift to the Collapse

The overall style drift of an article is measured as a distance in a standardized feature space. Each style feature is first z-scored against the pooled human-fake and V0 reference distribution, as in the left-hand side of Equation (23), which places every feature on a common scale regardless of its natural units. The style distance between an article $x_i = (x_{i,1}, \dots, x_{i,F})$ and a laundering article $l_i = (l_{i,1}, \dots, l_{i,F})$ is then the Euclidean norm of the difference between their z-scored feature vectors $x_i^z = (x_{i,1}^z, \dots, x_{i,F}^z)$ and $l_i^z = (l_{i,1}^z, \dots, l_{i,F}^z)$ as in Equations (23)–(25):

$$dist_{style}(x_i^z, l_i^z) = \sqrt{\sum_{f=1}^F (x_{i,f}^z - l_{i,f}^z)^2} \quad (23)$$

$$x_{i,f}^z = \frac{x_{i,f} - \mu_f^{(x)}}{\sigma_f^{(x)}}, f = 1, \dots, F; i = 1, \dots, M \quad (24)$$

$$l_{i,f}^z = \frac{l_{i,f} - \mu_f^{(l)}}{\sigma_f^{(l)}}, f = 1, \dots, F; i = 1, \dots, M \quad (25)$$

In Equations (24) and (25), F is the number of features; the terms $x_{i,f}$ and $l_{i,f}$ are the raw values of f -th feature in the x_i and l_i , respectively; while $\mu_f^{(x)}$ and $\mu_f^{(l)}$ are the mean, $\sigma_f^{(x)}$ and $\sigma_f^{(l)}$ are the standard deviation for f -th feature of the x_i and l_i , respectively. The anchor is the human-fake source article, so a larger style distance means the laundered article moved further from human-fake writing. The linguistic measurements connect to the detection results in two ways. First, the WF chain produces the largest style drift, with a style distance of about 4.1 against 2.4 for the CF chain, which matches the finding in Section 4.4 that WF is the strongest attack. Second, the features that laundering changes most are stylistic, while the content features stay stable, which matches the WELFake decomposition where about 37 points of the score were style. The mechanism behind the AI-style-shift vulnerability is therefore concrete and measurable. Laundering moves the text into the machine register, and any detector that learns the register loses exactly the signal it depended on.

6. Discussions

6.1. Why Do Standard Benchmarks Overstate Performance?

The main practical message of this study is that a near-perfect score on a standard fake news benchmark should not be read as near-perfect detection of fake news. On WELFake, about 37 AUROC points of the reported score were driven by writing registers rather than veracity. The laundering regime cannot reveal this, because it compares untouched human-real text against fake text, so register and veracity always move together. A detector is free to pick the easier of the two signals, and register is the easier one. The high benchmark number is real, but it answers a different question from the one the benchmark claims to ask.

6.2. Register-Controlled Evaluation as Standard Practice

We recommend that detection of fake news is reported with a register-controlled regime next to the standard one. The procedure is light. The real class is passed through the same rewriting transformation as the fake class, and the detector is then evaluated on the matched pair. The gap between the two regimes is an approximate measurement of how much the standard score depended on the register. A detector that holds its score under register control has earned its number. A detector that collapses, as the large models do on WELFake, has been reading style. Reporting both regimes turns a hidden confound into a visible and comparable quantity.

6.3. Implications for Deployment

The practical message of this study concerns how detectors are evaluated and selected rather than one-click evasion attacks. Under the laundering regime of Phase 1, where only the fake class is laundered, the detectors stay robust, so an adversary who merely paraphrases fake content does not break them. Fragility appears only once the register cue itself is removed, which is what the register-controlled regime of Phase 2 does. The risk is therefore an inflated benchmark score that leads to over-reliance and to the wrong choice of detector. A practitioner who reads a near-perfect standard number may deploy a detector that has learned writing register rather than veracity, and that detector will fail on

any future input where register and label come apart, for example as machine paraphrase becomes the norm for both true and false text. The capacity inversion sharpens the point. The instinct is to deploy the largest available model, yet on a style confounded benchmark the largest model is the one with the largest hidden register component. A detector should therefore be selected on its register-controlled score next to its standard score, not on the standard score alone.

6.4. Comparison with Prior Style and Laundering Attacks

Three studies sit closest to ours in the literature on attacks against fake news detectors. SheepDog of Wu et al. [26] are the nearest neighbors by setup because they also pass both the real class and the fake class through a commercial language model and they report the drop in F1 on a fine-tuned RoBERTa-base detector. They prompt GPT-3.5 to rewrite each article in the style of a named publisher, the real class is moved to a tabloid style such as the National Enquirer and the fake class is moved to a mainstream style such as the New York Times, and they report a drop of up to about 30 F1 points across PolitiFact, GossipCop and LUN, with the largest drop of about 30 points on LUN. AdStyle of Park et al. [27] use adversarial style prompts generated against a target detector and answers with a style invariant training scheme. Das and Dodge [28] use the exact term laundering and pass both classes through PEGASUS, GPT and Llama paraphraser on the COVID-19 and LIAR corpora, where the strongest paraphraser drops a BERT detector by about five F1 points. The TruthOverTricks study of Wan et al. [38] casts writing style as one of four shortcuts and reports accuracy drops of up to 80 points on sixteen corpora when those shortcuts are removed.

Two design differences separate those studies from ours and explain why the conclusions diverge. The first is the direction of the register manipulation. SheepDog and AdStyle swap the registers of the two classes through different prompts, the real class is moved to a tabloid style, and the fake class is moved to a mainstream style, so the two classes are still in different registers after the attack and the labels of those registers have flipped. We do not swap registers, we match them. Both the real seed and the fake seed of every article are laundered through the same paraphrase chains, so both classes end up in one register, and the detector must choose without the register cue. The second difference is the framing. SheepDog, AdStyle and Das and Dodge position laundering as an evasion attack and report a drop, TruthOverTricks positions writing style as one of several shortcuts in a broader catalogue, and all four works answer with a defense or a debiasing scheme. We use the same operation as a controlled lesion of the laundering regime, and we read the resulting collapse as a measurement of the size of the register confound, not as an attack to be patched.

On numbers the differences follow from the design. SheepDog reports RoBERTa-base F1 drops of about 11 points on PolitiFact, 3 points on GossipCop and 30 points on LUN. Our AUROC drop on WELFake reaches 37 points on average and three of the largest detectors fall to the chance line; on IFND, the drop is about 9 points, and on GossipCop, the drop is near zero, which matches their finding on the same corpus. Two patterns appear in our results that prior work does not report. The first is the capacity inversion, where the largest detectors collapse hardest under register matching, a pattern that prior style and laundering attacks do not separate. The second is that the residual AUROC on the register matched test is near 50% for the largest detectors on WELFake, because under that operation, the detector is not reading a softer veracity signal, it is reading nothing at all. Table 12 summarizes the comparison.

Table 12. Comparison of laundering and style attacks against transformer-based fake news detectors.

Work	Benchmarks	Detector	Laundering Scope	Framing	Headline Drop
SheepDog, Wu et al. 2024 [26]	PolitiFact, GossipCop, LUN	RoBERTa-base	Different prompt styles for both classes	Attack, defense by style invariant training	F1 drops up to 30 points
AdStyle, Park et al. 2025 [27]	three fake news corpora	RoBERTa	Adversarial style prompts on the fake class	Attack, defense by adversarial style training	F1 drop comparable to SheepDog
Das and Dodge 2025 [28]	COVID-19, LIAR	BERT, T5, Llama, classical	Same paraphraser on both classes	Laundering attack	BERT F1 drops about 5 points, COVID-19, PEGASUS
TruthOverTricks, Wan et al. 2025 [37]	sixteen misinformation corpora	BERT, DeBERTa, Mistral, Llama-3	Distribution shift on shortcut features	Shortcut catalogue, mitigation	Accuracy drops up to 80 points
Ours	WELFake, GossipCop, IFND	seven detectors, BERT-base to Longformer	Same V0 and paraphrase chain on both classes	Register-controlled lesion of the standard regime	AUROC drop 37 points WELFake, 9 IFND, near 0 GossipCop

Reproduction on the Shared and the New Benchmark

To anchor the comparison empirically, we re-implemented the SheepDog publisher restyle attack of Wu et al. [26] verbatim. We used their exact prompt, Rewrite the following article using the style of a named publisher, and their exact attacker, GPT-3.5, on the 400 articles test split of GossipCop and on the WELFake test split that prior work has not evaluated. The real class was rewritten in the style of The National Enquirer and the fake class in the style of The New York Times, matching their cross-style template. The restyled test sets were then scored with the same RoBERTa-large detectors used in our main results, which mean over the five training seeds.

On GossipCop, both manipulations stay close to the clean baseline. The SheepDog attack drops RoBERTa-large by about six AUROC points and seven F1 points, like the small drop they reported on the same corpus with RoBERTa-base. Our register matching produces no drop at all on this benchmark, because there is little residual register cue for either operation to destroy, which the pre-laundering style probe confirms with a GossipCop register margin of 0.597, small but above chance. The two pipelines agree that GossipCop sits at the low end of the style reliance spectrum.

On WELFake, the result is more revealing, and it is the one that supports the methodological distinction we drew in Section 6.4. The same SheepDog attack moves RoBERTa-large on WELFake from a clean AUROC of 99.9% to 98.0%, a drop of less than two points. Our register matching moves the same detector from 99.9% to 59.2% AUROC under the Phase 2 mean across all eighteen cells, and to 49.4% under the WF R3 chain, which is the chance line. The 38.8 points gap under the Phase 2 mean, and the 48.7 points gap under the strongest chain, are direct empirical measurements of how much register swapping and register matching differ as test time operations on the same detector and the same test split. Swapping the registers of the two classes through different prompts moves the inputs but leaves the register cue available, only with the labels of the two registers flipped, so a fine-tuned detector adapts at the chunk level and the score barely shifts. Matching the

registers through a common transformation removes the cue, and the detector has nothing left to read. Table 13 reports the head-to-head numbers per benchmark.

Table 13. SheepDog publisher restyle attack and our register-controlled lesion on the RoBERTa-large detector, mean over five seeds.

Benchmark	Regime	AUROC (%)	Accuracy (%)
WELFake	Phase 1 baseline (HR vs. HF)	99.87	97.90
WELFake	SheepDog publisher attack	98.03	92.30
WELFake	Phase 2 mean	59.21	50.06
WELFake	Phase 2 WF R3	49.37	50.11
GossipCop	Phase 1 baseline (HR vs. HF)	88.24	81.50
GossipCop	SheepDog publisher attack	82.11	72.15
GossipCop	Phase 2 mean	88.84	81.53
GossipCop	Phase 2 WF R3	90.07	82.22

7. Limitations and Threats to Validity

Several limitations should be kept in mind when reading those results. This study covers three datasets and the English language only, so the graded spectrum of style reliance may look different on other corpora and other languages. The factuality control depends on two large language model judges, and although the preserved subset is the conservative intersection of their verdicts, the judges are not a perfect oracle. The detectors were trained in balanced samples of 2000 items, which keep the datasets comparable but are smaller than some full benchmarks. The learning curve in Section 4.3 shows the effect is stable from 500 to 8000 items, so the sample size is not the limiting factor, but the scope of three English corpora is, and we mark wider coverage as future work. The training-size learning curve was run on WELFake only, owing to the compute and time budget of the revision. WELFake was selected as the corpus where the register margin is largest and the capacity inversion strongest, so it is the most demanding test of the small-sample concern, and we leave replication across the other corpora to future work. Finally, the study evaluates trained discriminative detectors. Whether a large language model used directly as a zero-shot detector shows the same collapse is an open question, which we leave to future work. To separate style removal from a generic distribution shift, we report an adversarial training control in Section 4.6; extending that control across corpora and to other retraining regimes is a natural next step.

8. Conclusions

This paper showed that the high accuracy of fake news detectors on standard benchmarks is, in large part, a measurement of writing register and not of veracity. We named this failure the AI-style-shift vulnerability, and we gave a register-controlled evaluation regime that exposes it. The real class is laundered through the same chains as the fake class; thus, both classes are read in one machine register, and the register cue is removed. Under that regime, detection on WELFake collapses from about 99% to about 62% AUROC, larger detectors collapse harder, and the largest neural models approach chance. The effect is benchmark dependent. It is large on WELFake, mild on IFND and near zero on GossipCop, which shows that the size of the inflation depends on how each benchmark's fake class was built rather than being a single uniform effect. The collapse is graded across benchmarks and is statistically confirmed. Based on those results, we recommend that register-controlled evaluation is reported as a standard practice in the field. Future work should extend the regime to more languages and domains, test large language models used directly as detectors, and study whether adversarial training on laundered text can recover

genuine veracity signals. The vulnerability we document is specific to text-only detectors. Signals outside the text cannot be laundered by paraphrasing, because rewriting an article changes neither who shared it nor where it started. Hybrid detectors that combine text with the social graph or with other non-textual signals therefore offer structural robustness to the style-based attack studied here. Multimodal fake news detectors make this concrete. EANN [39] and MVAE [40] classify a post by pairing its text with the attached image and DEFEND [41] pairs the article with the user comments it attracts. A paraphrase chain rewrites the article text and leaves the image and the comment thread untouched, so whatever part of such a detector's decision rests on those signals is out of reach of the operation studied here. A related line of work classifies users rather than articles: Corizzo and colleagues [42] combine the topology of the interaction network with the semantics of the content users share, and Benedetti and colleagues [43] add the spatial information of users as a third perspective. Those methods target accounts rather than articles, so they do not replace content-level veracity assessment, but they share the same structural property. We mark hybrid detection as a direction for future work.

Author Contributions: Conceptualization, J.M. and R.A.; methodology, J.M. and R.A.; validation, J.M. and R.A.; data curation, J.M.; writing—original draft preparation, J.M.; writing—review and editing, J.M. and R.A.; funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research by Prof. Dr. Ramiz Aliguliyev was supported by Azerbaijan Technical University, Grant No. AzTU-DQL-2025-M01/60182631.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code for the cross-LLM paraphrase laundering pipeline, the configuration files, and a representative sample of the laundered corpus are openly available at <https://github.com/mehdiyevcs/cross-llm-paraphrase-laundering> (accessed on 30 July 2026). The full laundered corpus, the preserved subset, and the trained detectors are available from the corresponding author on reasonable request or can be regenerated from the source benchmarks using the released pipeline so that the register-controlled evaluation can be reproduced. The source benchmarks WELFake, GossipCop, and IFND are publicly available from their original providers and are not redistributed here.

Acknowledgments: During the preparation of this manuscript the authors used a large language model (Claude Opus 5 Model, Anthropic) to assist with language editing, rewording of drafted paragraphs, and general readability improvement of sections originally drafted by the authors. The tool was not used to design the study, generate or alter research data or results, select references, or draw any conclusions. All scientific content, analysis, argumentation and conclusions are the authors' own. After using the tool, the authors reviewed and edited the material and took full responsibility for the final text of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Verma, P.K.; Agrawal, P.; Amorim, I.; Prodan, R. WELFake: Word Embedding over Linguistic Features for Fake News Detection. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 881–893. [CrossRef]
2. Kaliyar, R.K.; Goswami, A.; Narang, P. FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach. *Multimed. Tools Appl.* **2021**, *80*, 11765–11788. [CrossRef] [PubMed]
3. Nasir, J.A.; Khan, O.S.; Varlamis, I. Fake News Detection: A Hybrid CNN-RNN Based Deep Learning Approach. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100007. [CrossRef]
4. Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut Learning in Deep Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]

5. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic Detection of Fake News. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3391–3401.
6. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Toronto, ON, Canada, 3–10 March 2021; pp. 610–623.
7. Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; Iyyer, M. Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 27469–27500.
8. Schuster, T.; Schuster, R.; Shah, D.J.; Barzilay, R. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Comput. Linguist.* **2020**, *46*, 499–510. [[CrossRef](#)]
9. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 231–240.
10. Zhou, X.; Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods and Opportunities. *ACM Comput. Surv.* **2020**, *53*, 109.
11. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [[CrossRef](#)]
12. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
13. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [[CrossRef](#)] [[PubMed](#)]
14. Sharma, D.K.; Garg, S. IFND: A Benchmark Dataset for Fake News Detection. *Complex Intell. Syst.* **2023**, *9*, 2843–2863. [[CrossRef](#)] [[PubMed](#)]
15. Wang, W.Y. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 422–426.
16. Ribeiro, M.T.; Wu, T.; Guestrin, C.; Singh, S. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4902–4912.
17. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text Be Reliably Detected? *arXiv* **2023**, arXiv:2303.11156.
18. Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.R.; Smith, N.A. Annotation Artifacts in Natural Language Inference Data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 107–112.
19. McCoy, T.; Pavlick, E.; Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3428–3448.
20. Niven, T.; Kao, H.-Y. Probing Neural Network Comprehension of Natural Language Arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4658–4664.
21. Gehrmann, S.; Strobel, H.; Rush, A.M. GLTR: Statistical Detection and Visualization of Generated Text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, System Demonstrations, Florence, Italy, 29–31 July 2019; pp. 111–116.
22. Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C.D.; Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 24950–24962.
23. Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Wang, J. Release Strategies and the Social Impacts of Language Models. *arXiv* **2019**, arXiv:1908.09203.
24. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending Against Neural Fake News. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 9054–9065.
25. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Wu, Y. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597.
26. Wu, J.; Guo, J.; Hooi, B. Fake News in Sheep’s Clothing: Robust Fake News Detection against LLM-Empowered Style Attacks. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 3367–3378.

27. Park, S.; Han, S.; Xie, X.; Lee, J.-G.; Cha, M. Adversarial Style Augmentation via Large Language Model for Robust Fake News Detection. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 4024–4033.
28. Das, R.K.; Dodge, J. Fake News Detection after LLM Laundering: Measurement and Explanation. *arXiv* **2025**, arXiv:2501.18649.
29. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* **2019**, arXiv:1910.01108.
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
31. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150.
32. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Rush, A.M. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
33. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall: New York, NY, USA, 1993.
34. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [[CrossRef](#)]
35. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python. *Zenodo* **2020**. [[CrossRef](#)]
36. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
37. Horne, B.D.; Adali, S. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. In Proceedings of the 11th International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; pp. 759–766.
38. Wan, H.; Wu, J.; Luo, M.; Zeng, Z.; Su, Z. Truth over Tricks: Measuring and Mitigating Shortcut Learning in Misinformation Detection. In Proceedings of the 39th Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025; pp. 1–15.
39. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2018; pp. 849–857.
40. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of the World Wide Web Conference*; ACM: New York, NY, USA, 2019; pp. 2915–2921.
41. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2019; pp. 395–405.
42. Corizzo, R.; Pio, G.; Barracchia, E.P.; Pellicani, A.; Japkowicz, N.; Ceci, M. HURI: Hybrid User Risk Identification in Social Networks. *World Wide Web* **2023**, *26*, 3409–3439. [[CrossRef](#)]
43. Benedetti, F.; Pellicani, A.; Pio, G.; Ceci, M. IMMENSE: Inductive Multi-perspective User Classification in Social Networks. *Online Soc. Netw. Media* **2025**, *50*, 100335. [[CrossRef](#)]

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