

A Rewrite That Keeps the Label Has Already Read It: Why an LLM Paraphrase Can Raise a Sentiment Classifier's Accuracy Without the Classifier Getting Better, Four Accounts of the Gain, and the Re-Annotation and Neutral-Rate Controls That Would Separate Them

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A 2026 study reports that sentiment classifiers score higher on reviews rewritten by small language models than on the human originals, by up to 5.8 accuracy points, and reads the gain as stylistic alignment: the rewrite strips idiosyncrasy and hands the classifier text closer to what it was trained on. It concludes that benchmarks cleaned or paraphrased by language models will overstate classifier performance. Two older literatures point the other way. Paraphrase and style-transfer attacks drive the same classifiers down, and a 2025 study of fifty thousand rephrased tweets finds that rewriting pulls sentiment toward neutral, a pull its robustness checks found again when the prompt asked for the sentiment to be kept. In this analysis the author reviews and synthesises published results to show that neither direction measures the classifier alone, because every one of these designs scores the rewritten text against the label of the text it replaced. The paper separates four accounts of a rewrite-induced gain: alignment of style with the classifier's training data; a fall in the rate of neutral predictions, which the anchor's scoring counts as errors; the paraphraser acting as an annotator that decides the polarity and then writes it plainly; and the removal of hard, mixed and ironic items. It shows from the anchor's released code that its prompt demands one of two polarities and that a neutral prediction is always scored wrong, that its own count of label flips is mislabelled: the printed 291 is the net gain, not the number of improvements, and that a test-time rewriting study found rewritten inputs still sitting with the out-of-distribution data in embedding space. The paper consolidates the measured values in two tables and a figure, states a licensing procedure as an algorithm, counts from the anchor's released predictions that a fall in neutral predictions carries about half of the largest gain, and names the re-annotation, paraphraser-baseline and prompt-ablation controls that would settle which account holds for the rest.

Keywords: sentiment analysis; AI-generated text; paraphrase robustness; label preservation; distribution shift; evaluation methodology; selective classification; sarcasm

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Sentiment classifiers are among the oldest deployed tools in language processing, and the text they are applied to is changing under them. Population-level estimates find that a growing share of ordinary written text is now produced or substantially edited by language models: Liang et al. (2025) estimate that by late 2024 roughly 18 percent of the text in a large corpus of financial consumer complaints shows signs of model assistance, and an earlier estimate by Liang et al. (2024) put the share of substantially model-modified text in AI conference peer reviews between 6.5 and 16.9 percent. Machine paraphrases are hard for people to tell from human writing; in a study of 105 participants by Wahle et al. (2022), participants identified GPT-3 paraphrases at a mean accuracy of 53 percent. If a sentiment classifier behaves differently on text that a model has touched, every measurement made with it on post-2022 text carries a term nobody is reporting.

In August 2026, Shroff (2026) measured that term directly and found it had the opposite sign from the one most readers would expect. Five thousand Yelp reviews

were rewritten by two small instruction-tuned models, Qwen3.5-4B and Gemma4-E4B, and two standard classifiers, the lexicon-based VADER (Hutto and Gilbert, 2014) and a RoBERTa model further trained on tweets, were scored on the originals and on the rewrites. The RoBERTa classifier was more accurate on both sets of rewrites, by 5.8 points for one paraphraser and 3.7 for the other; VADER gained 3.2 points on one and lost 1.7 on the other. The paper names the effect cross-domain stylistic alignment, proposes that the rewrites strip reviewer idiosyncrasy and so move the text toward the short, plain register the Twitter-trained classifier knows, and draws a methodological conclusion: a sentiment benchmark that has been augmented, cleaned or paraphrased by a language model will overstate classifier performance relative to organic text, so the generative source of each item should be recorded.

The result lands between two bodies of work that have long pointed in opposite directions. The adversarial-robustness literature established that meaning-preserving rewrites break classifiers: syntactically controlled paraphrases fool pretrained models (Iyyer et al., 2018), word-substitution attacks drive accuracy down sharply (Jin et al., 2019), and rewrites that change only the style

of a sentence reach attack success rates above 90 percent on SST-2 sentiment against every model tested (Qi et al., 2021). A newer line on model-rewritten user content finds that rewriting moves sentiment in a direction that should cost a classifier accuracy rather than add it: Wang et al. (2025) rephrased 50,000 climate-change tweets with a GPT model and found the mean VADER positive and negative scores both fell, toward neutral; in a robustness check on 1,000 of the tweets, the shift persisted when the prompt explicitly asked the model to keep the tweet's sentiment unchanged. And a third line, on test-time augmentation, sides with the anchor: O'Brien et al. (2024) rewrote out-of-distribution test inputs with a language model and raised a BERT sentiment classifier's accuracy on them by 5.12 points on average.

So within a few years the field holds published results in which model rewriting lowers classifier accuracy, results in which it raises it, and results in which it changes the sentiment the text carries. This paper asks what a rewrite-induced change in classifier accuracy measures, and what would have to be controlled before it could be read as a property of the classifier. In this analysis the author reviews and synthesises published results to show that the answer turns on a design choice all of these studies share and none of them tests: the rewritten text is scored against the label of the text it replaced. The answer comes in four parts.

First, a change in accuracy under rewriting is a sum of two terms: how well the classifier reads the new text, and how far the new text still deserves the old label. Every study located scores only the sum. Where the rewrite is selected to defeat the classifier, label change inflates the measured loss; where the rewrite is instructed to preserve a stated polarity, it can inflate a measured gain.

Second, the anchor's gain fits at least four accounts: stylistic alignment, as proposed; a fall in the rate at which a three-class classifier predicts neutral, which the anchor's scoring counts as wrong on two-class labels; the paraphraser acting as an annotator that decides the polarity and writes it out; and the removal of the mixed, hedged and ironic items on which classifiers fail. The released code and the anchor's own numbers bear on each.

Third, the published record does not favour alignment over the other three. A test-time rewriting study that found large gains also found that the rewritten inputs stayed in the out-of-distribution region of embedding space, and a classifier fitted on paraphrased text gained slightly even though no shift between training and test text remained to be aligned.

Fourth, the anchor's released predictions already give the neutral-prediction rate and polar-only accuracy, which split the largest gain about evenly between the two, and three further controls would separate the accounts for the polar part: blind re-annotation of a sample of rewrites, using the paraphraser's own zero-shot label as a

baseline, and ablating the polarity instruction in the rewriting prompt.

The paper's contribution is the decomposition, a second reading of the anchor's tables and released code that narrows what they establish, a reconciliation of the opposite-signed literatures along the label-inheritance axis, and a licensing procedure. It does not argue that model rewriting leaves classifiers unaffected, or that the anchor's methodological warning is wrong in kind. It argues that the warning's direction, "systematically overestimate", is not licensed by the evidence, and says what would license a direction.

1.1 Relation to the author's earlier papers

Four of the author's papers sit near this one. An earlier published review examined the detection of AI-generated text under paraphrasing; it concerned whether a detector can tell that text was machine-written, which is a different question from whether a classifier of some other property reads machine-touched text correctly, and it did not treat sentiment. A second paper argued that a score obtained without one input modality is not a measure of reliance on that modality; it shares with this one the habit of asking what a contrast between two input conditions licenses, but it concerned audio-language benchmarks. A third reconciled two jailbreak studies whose degradation curves had opposite signs by listing accounts that fit the same data; Section 6 below reconciles opposite-signed rewriting results by a different route, through what each design holds fixed about the label. A fourth argued that an effect measured in agent scaffolds is a contrast conditional on what was held fixed; the point carries over in spirit, and Section 3 makes the corresponding statement for rewriting precise. None of the four treats sentiment classification, label preservation, or model-rewritten text as a test condition.

2. Scope, and what this paper does not claim

This paper analyses one kind of measurement: a fixed text classifier is scored on a set of human-written items and on model-rewritten versions of the same items, and the difference in accuracy is attributed to the rewriting. The anchor is sentiment classification, and most of the evidence used is from sentiment, but the decomposition in Section 3 applies to any label inherited across a transformation, including toxicity and topic. Detection of machine-written text is out of scope except where it bears on how rewritten text differs from human text.

The anchor's numbers are taken as printed. Where this paper reports what the anchor's released notebook does, it describes the code as published with the paper, and says so. Counts from the anchor's released per-item predictions are the author's tabulation and are marked. Counts and differences computed from printed values are the author's arithmetic and are marked when they appear. The worked example in Section 3.3, which shows how much a change in the neutral-prediction rate could move accuracy, is a derivation under stated assumptions, not an estimate of what happened in the

anchor; it says so where it appears.

The paper does not claim that stylistic alignment is false. Section 5.1 sets out the evidence for it, which is real. When this paper says a quantity is not known, it means that no source found in the search for this paper reports it; it is a claim about the published record, not about what is knowable.

3. What an accuracy change under rewriting is made of

3.1 Two labels, one score

Let x be a human-written item with gold label y , assigned by whatever process built the benchmark. Let T be a rewriting procedure, a language model with a prompt and decoding settings, and let f be the classifier. The quantity every study in this paper reports is the difference

$$D = P[f(T(x)) = y] - P[f(x) = y],$$

the change in agreement with the original label when the classifier reads the rewrite instead of the original.

The claim the anchor attaches to D is about f : that the classifier "relaxes" on rewritten text, reading it more accurately. That claim concerns a different quantity. Let y_T be the label the same annotation process would assign to $T(x)$ if it were annotated afresh, blind to x . The classifier's accuracy on the rewritten text is $P[f(T(x)) = y_T]$, and the claim about the classifier is a claim about

$$D_{\text{own}} = P[f(T(x)) = y_T] - P[f(x) = y].$$

The two differ by a term that is non-zero only on items whose label the rewrite changed:

$$D - D_{\text{own}} = P[f(T(x)) = y, y_T \neq y] - P[f(T(x)) = y_T, y_T \neq y].$$

The identity follows from the definitions of D and D_{own} . It says that D measures the classifier only when the rewriting preserves labels, and that on the items where it does not, the inherited scoring rewards a classifier for disagreeing with the text in front of it. None of the studies of rewritten text located for this paper reports y_T , so none reports D_{own} ; each reports D .

3.2 Which way the label-change term points depends on how the rewrite was chosen

The sign of the label-change term is not fixed. It depends on how the rewrites were produced and selected.

In an adversarial design, candidate rewrites are searched for one the classifier gets wrong. If some candidates change the label, the search favours them, because a rewrite whose meaning has flipped is an easy way to make the classifier disagree with the inherited label. The measured loss then includes a contribution from label change that the classifier is not responsible for. Section 6 shows that the human validation studies of adversarial text found exactly this, at large rates.

In an instructed design, the rewriting model is told what to keep. If it is told to keep a polarity, it must first decide what the polarity is, and the rewrite it produces will tend to state that polarity plainly. On items whose original was ambiguous, mixed or ironic, the rewrite can therefore carry a clearer version of the label than the original did, or a clearer version of the wrong label if the paraphraser misread it. The first case raises D without any change in what the classifier can do; the second lowers it.

In an unselected, uninstructed design, a plain paraphrase prompt with no label mentioned, the term has no designed direction, and its sign is whatever the rewriting model's habits produce. Section 5.2 shows that the best-documented habit, at least for GPT models rephrasing tweets, is to pull sentiment toward neutral.

3.3 A three-class reader on two-class labels

A second term is specific to the anchor's design and is mechanical. The anchor describes its RoBERTa classifier as the Cardiff NLP twitter-roberta-base-sentiment-latest model, fine-tuned on TweetEval sentiment data. That model descends from RoBERTa-base (Liu et al., 2019) through Cardiff NLP's Twitter-adapted continuations of it (Loureiro et al., 2022), and the TweetEval sentiment task it was fine-tuned for asks whether a tweet is positive, negative or neutral, using the SemEval-2017 Twitter data (Barbieri et al., 2020; Rosenthal et al., 2017). Yelp Polarity, built by Zhang et al. (2015), has two labels. VADER, as the anchor's notebook applies it, also returns neutral when the compound score lies between -0.05 and 0.05.

When a three-class classifier is scored against two-class gold labels, every neutral prediction is an error. Write n for the fraction of items the classifier calls neutral, and a for its accuracy on the items it calls positive or negative. Then accuracy is $(1 - n)a$, and a rewrite can raise accuracy by lowering n with no change in a . As a worked illustration only, and not as an estimate of the anchor's values: at a fixed polar accuracy of 0.90, a neutral rate of about 8.7 percent gives an accuracy of 0.822, and a neutral rate of about 2.2 percent gives 0.880. A fall of about six and a half points in the share of neutral predictions would by itself reproduce the anchor's 82.2-to-88.0 change for RoBERTa on the Qwen rewrites. The anchor does not print its neutral rates, but its released per-item predictions contain them. By the author's count from those predictions, RoBERTa called 436 of the 5,000 originals neutral (8.72 percent) and 276 of the Qwen rewrites (5.52 percent), and its accuracy on the items it called positive or negative rose from 90.07 to 93.18 percent. Putting those values into $\text{accuracy} = (1 - n)a$, the drop in n is worth about 2.9 of the 5.8 points and the rise in a about the other 2.9. So the neutral-rate term is real and carries about half the RoBERTa gain on the Qwen rewrites; the polar half is left for the alignment, annotator and hard-item accounts of Section 5 to explain.

4. What the anchor measured

4.1 The design

Shroff (2026) reports three experiments. In the first, the RoBERTa classifier's confidence was compared on 867 tweets their authors had labelled sarcastic and 2,601 they had not, from the English training set of the SemEval-2022 iSarcasmEval task (Abu Farha et al., 2022); mean confidence was 0.741 on the sarcastic tweets and 0.770 on the rest, a difference the paper reports as significant by a Mann-Whitney test at $p = 2 \times 10^{-6}$. iSarcasmEval carries no sentiment polarity labels, so this experiment measures confidence, not accuracy. In the second, 5,000 Yelp Polarity reviews, balanced across the two labels, were rewritten by Qwen3.5-4B and Gemma4-E4B at temperature zero, and both classifiers were scored on all three versions. In the third, an abstention rule withheld predictions on items where RoBERTa's top-class probability was below 0.6, and a comparison of two uncertainty signals on 300 iSarcasmEval items was run with ten samples per item from Qwen3.5-4B.

The released notebook fixes several details the text leaves open. The rewriting prompt asks the model to rewrite the review in its own words and to "Preserve the exact sentiment (positive or negative) and key points". Accuracy is computed by comparing each classifier's label, drawn from positive, negative and neutral, against the gold label mapped to "negative" or "positive", so a neutral prediction can never be correct. RoBERTa inputs are truncated at 512 tokens. If a batch fails at inference and the item-by-item retry also fails, the item is assigned the label neutral with a score of 0.34, which is scored as an error.

4.2 The accuracy table

Table 1 collects the values from the anchor on which this paper relies, with the location in the anchor named on each row.

4.3 The flip count is mislabelled

The anchor checks its headline by counting label changes. For RoBERTa on the Qwen rewrites it reports 573 items whose predicted label differs between original and rewrite, of which 291 are described as net improvements and 282 as net losses, and it calls the split "near-symmetric but slightly positive". Read as items that went from wrong to right and items that went from right to wrong, those counts imply a net gain of nine items out of 5,000, about 0.18 points. The headline is 5.8 points, which on 5,000 items is about 290 items. The two cannot both be right under that reading.

The labels are wrong, not the count. The anchor's released per-item predictions reproduce its printed accuracies and its 573 changed items, and they give the transition table the text omits. By the author's count, 386 of the changes went from wrong to right, 95 from right to wrong, and 92 between two wrong labels, for instance from neutral to the wrong polarity. The printed 291 is the net, 386 minus 95, which is 5.82 points on 5,000

items and matches the headline; the printed 282 is 573 minus 291 and counts no particular kind of change. The changes behind the gain run about four to one in the label's favour, not near-symmetrically, so the sentence that presents the split cannot be used as the check it is offered as until its labels are corrected.

4.4 The VADER rows

VADER is a fixed, human-curated lexicon combined with hand-written rules (Hutto and Gilbert, 2014); it has no training distribution in the sense the alignment account needs. Its accuracy rose by 3.2 points on the Qwen rewrites and fell by 1.7 on the Gemma rewrites. The anchor says the rewrites "mostly improve" VADER accuracy. Two of two RoBERTa rows rose and one of two VADER rows rose. A lexicon cannot be aligned with by stylistic convergence, so whatever moved VADER's accuracy is a change in the words the rewrite chose, and it moved in different directions for the two paraphrasers. That is direct evidence, inside the anchor, that the rewriting model's habits, not a general property of AI text, set the sign.

4.5 The abstention numbers

The abstention rule flagged 700 of 5,000 human-written Yelp reviews, and accuracy on the remaining 4,300 was 88.9 percent against 82.2 percent on all items. The anchor states that the 0.6 threshold was chosen as a trough in the confidence distribution of the same data and acknowledges this may overstate the gain. Accuracy on the flagged items was 41.3 percent, which the anchor describes as barely above chance for a three-class problem. The gold labels are two-class, so the reference point for chance is 50 percent, and the flagged items scored below it. A classifier can score below binary chance on binary labels only if it is systematically wrong or if some of its predictions cannot be right; the neutral predictions are of the second kind. The anchor does not report the composition of the flagged set, but its released predictions do. By the author's count, 269 of the 700 flagged items are neutral predictions, and accuracy on the other 431 is 67.1 percent, so part of what the abstention rule removes is the scoring mismatch rather than uncertain reading.

5. Four accounts of the gain

Each account below is stated with the evidence for it, the evidence against it, and what it predicts that the others do not. Table 3 in Section 10 collects the predictions.

5.1 Stylistic alignment

The anchor's account is that the classifier was trained on short informal tweets, Yelp reviews are longer and more idiosyncratic, and the rewrite moves them toward the classifier's home register. The account has real support. Pretrained Transformers lose accuracy under shifts between review sources, as Hendrycks et al. (2020) showed for SST-2 and IMDb, though the same study found only minor drops across Yelp's cuisine-based splits. The BOSS benchmark of Yuan et al. (2023) was built to

Table 1. Values reported by Shroff (2026), with the anchor table or section named on each row. Entries in the last column marked "author" are the author's arithmetic on printed values or the author's tabulation of the anchor's released per-item predictions.

Row	Anchor location	Condition	Reported value	Bearing
1	Table 2	RoBERTa, Qwen3.5-4B	82.2 to 88.0 percent (+ 5.8)	headline gain
2	Table 2	RoBERTa, Gemma4-E4B	82.2 to 86.0 percent (+ 3.7)	86.0 is rounded from 85.96 in the released results; the + 3.7 is correct
3	Table 2	VADER, Qwen / Gemma	+ 3.2 / -1.7 points	sign differs by paraphraser
4	Sec. 4.2	label changes, RoBERTa-Qwen	573, split 291 / 282	291 is the net of 386 up and 95 down (author)
5	Sec. 4.2	label-change rate, all four	11.5 to 17.7 percent	one in six to one in nine (author)
6	Table 3	ECE, human / Qwen / Gemma	0.056 / 0.045 / 0.029	lower on rewrites
7	Table 4	abstention below 0.6	700 flagged; 82.2 to 88.9 percent	threshold set on test data
8	Table 4	flagged items only	41.3 percent accuracy	below the binary 50 percent
9	Table 5	entropy vs disagreement	AUROC 0.650 vs 0.646	proxy correctness label

make such shifts hard and found that standard robustness methods help little over plain fine-tuning. Against that background, O'Brien et al. (2024) proposed rewriting each out-of-distribution test input several times with a language model, prompted with in-distribution examples, and averaging the classifier's predictions over the rewrites and the original. On three-way sentiment, with Amazon reviews as the in-distribution set and DynaSent (Potts et al., 2021), SST-5 and a SemEval tweet set as the shifted sets, all taken from BOSS, this in-context rewriting raised BERT's out-of-distribution accuracy by 5.12 points on average while leaving in-distribution accuracy roughly unchanged or better.

That is a measured gain of the same size as the anchor's, under a design whose stated purpose is alignment. But the same paper tested whether alignment was what happened, and found it was not. Its appendix projects the embeddings of the in-distribution evaluation set, the out-of-distribution SST-5 inputs, and the rewrites, and reports that the rewrites remain largely in the same region as the out-of-distribution inputs they came from; the authors conclude that the rewriting does not bridge the domain gap even where it improves accuracy. The strongest published case for rewriting as domain adaptation thus comes with its own evidence that the gain does not run through the domain.

A second test comes from a study that was not about classifiers' robustness at all. Audinet de Pieuchon et al. (2024) asked whether language models can strip sentiment from Amazon reviews while keeping their topic, and as a baseline asked GPT-4 and Mistral-7B simply to paraphrase each review without changing its meaning. They then fitted a logistic-regression classifier on DistilBERT embeddings of the paraphrased reviews and tested it on held-out paraphrased reviews, so training and test text came from the same rewriting process. Sentiment accuracy was 0.885 on the originals, 0.891 on the Mistral paraphrases and 0.899 on the GPT-4 paraphrases, with intervals of about plus or minus 0.035 on each; the authors suggest that paraphrasing may amplify the signal already in the text. The differences lie inside the reported intervals and should

not be read as established. What matters for this paper is the direction and the design: in a setting where no gap between training and test domains existed to be closed, paraphrased text was at least as classifiable as the original, which is what the fourth account below predicts and what the alignment account does not.

Alignment predicts that the gain should shrink or vanish for a classifier trained on the target domain, and should not depend on what the prompt says about sentiment. Neither prediction has been tested on the anchor's data.

5.2 A fall in the neutral rate

The second account is the scoring mechanism of Section 3.3. If the rewrites carry less neutral-reading text than the originals, a three-class classifier predicts neutral less often, and accuracy against two-class labels rises.

There are two reasons to expect the anchor's rewrites to do this, and one strong reason to expect rewrites in general to do the opposite. The first reason is the prompt: it names the two admissible polarities, "positive or negative", and tells the model to preserve the one the review has. A model following that instruction on a review that hedges, praises the food and damns the service, or ends on a shrug, has to decide and has no third option to preserve. The second reason is that a Yelp Polarity review has no neutral option by construction, and its positive class, as its builders describe it, reaches into the middle of the rating scale. Zhang et al. (2015) describe the polarity labels as "considering stars 1 and 2 negative, and 3 and 4 positive". On that description, three-star reviews, the ones most likely to praise one thing and fault another, carry the positive label. Those are reviews a three-class tweet classifier can reasonably call neutral, and ones on which an instruction to preserve one of two polarities has much to change. They are not the whole of it: by the author's count from the anchor's released predictions, 284 of RoBERTa's 436 neutral calls on the originals fall on reviews labelled negative and only 152 on reviews labelled positive, a split this argument does not predict.

What a classifier means by neutral is itself a property of its training data. The three-way sentiment benchmarks disagree about where neutral ends: in building DynaSent, Potts et al. (2021) present evidence that their neutral class is more coherent than the comparable class in other three-way sentiment benchmarks, which is to say that the others' neutral classes are looser. A classifier fine-tuned on a looser neutral class will place more of a mixed review's probability mass there, and the neutral-rate term in Section 3.3 inherits that looseness. The size of the term therefore depends on which three-class classifier is used, not only on the rewrite.

The reason to expect the opposite is that, without such an instruction, the best-measured habit of model rewriting is to pull sentiment toward neutral. Wang et al. (2025) rephrased 50,000 climate-change tweets with text-davinci-003 and scored originals and rephrasings with VADER. Mean positive score fell from 0.126 to 0.095 and mean negative score from 0.123 to 0.083, both significant, while the compound score rose slightly, from 0.007 to 0.040; the dampening was larger for negative content. In robustness checks on 1,000 tweets, the shift toward neutral held under a prompt that asked for the sentiment to be kept unchanged, under a grammar-correction prompt, and with newer models up to GPT-4o-mini. On 10,000 long Amazon video-game reviews the same direction appeared at a much smaller size, about 0.003 on both the positive and negative scores. The pull toward neutral is a property of rewriting, not of model text in general: when a model writes its own reply rather than rephrasing someone else's text, the skew can run strongly positive, as in Pyreddy and Zaman (2025), where about 90 percent of ChatGPT's replies to tweets about generative AI were positive against about 27 percent of the human tweets. Translation shows the same pull as rewriting: Mohammad et al. (2016), who had Arabic social-media posts and their English translations annotated independently for sentiment, report that both manual and automatic translation often turn positive and negative texts into neutral ones.

Under the anchor's scoring, a rewrite that pulls sentiment toward neutral should lower accuracy, not raise it. The VADER result on the Gemma rewrites, a 1.7-point fall, is not neutralisation. By the author's count from the anchor's released predictions, VADER's neutral calls fell on those rewrites, from 103 to 89, and the loss comes from 450 changes from negative to positive against 272 from positive to negative: a positive shift in VADER's polar calls, which fits the small rise in the compound score that Wang et al. (2025) measured. The RoBERTa results on both paraphraser and the VADER result on Qwen are what the instructed-polarisation reading predicts. That the sign differs across paraphraser and classifier within one study is awkward for an account that attributes the effect to AI text as such.

This account predicts that accuracy restricted to items the classifier calls positive or negative should change

little, that the neutral-prediction rate should fall on the rewrites, and that removing the polarity instruction from the prompt should shrink the gain or reverse it. The anchor's released predictions answer the first two. By the author's count, RoBERTa's neutral rate fell from 8.72 percent on the originals to 5.52 percent on the Qwen rewrites and 6.42 percent on the Gemma rewrites, and its polar accuracy rose from 90.07 to 93.18 and 91.86 percent; VADER's neutral rate was small throughout, at 2.06, 1.80 and 1.78 percent. The neutral rate fell, as predicted, but polar accuracy did not stay flat: the neutral-rate term carries about half the RoBERTa gain on the Qwen rewrites and polar accuracy the rest, which leaves the alignment, annotator and hard-item accounts to explain the polar part. The prompt ablation has not been reported.

5.3 The paraphraser as annotator

The third account starts from what the instruction asks the rewriting model to do. To preserve a review's polarity, the model has to determine it. Instruction-tuned language models are strong zero-shot sentiment classifiers on reviews of this kind: in the evaluation of Zhang et al. (2023), on 500 sampled Yelp-2 test items, zero-shot accuracy was 92.2 percent for Flan-T5, 93.2 for text-davinci-003, 97.8 for ChatGPT and 98.2 for Flan-UL2, against 96.33 for a fully fine-tuned T5-large. Wang et al. (2023) report that ChatGPT handles negation and speculation, the polarity-shifting constructions that trip lexical and fine-tuned classifiers, better than a fine-tuned BERT. The advantage is not general: across 25 mostly subjective tasks, Kocon et al. (2023) found ChatGPT's quality about 25 percent below the best specialised systems on average in zero- and few-shot use. Binary review polarity is among the easiest of such tasks, and it is the one the anchor's paraphraser was asked to perform implicitly. The anchor's paraphraser is far smaller than the models above, and their zero-shot accuracy on Yelp is not reported, so the size of the effect in the anchor is not known. The mechanism does not need the paraphraser to be excellent, only better than the classifier on the items where the classifier fails.

On those items, a polarity-preserving rewrite does two things at once. It reads the review, which the classifier failed to do, and it writes the result in words the classifier can read. The classifier then gets the item right, and D records a gain that belongs to the paraphraser. This is the same structure that Gururangan et al. (2018) identified in natural language inference: crowdworkers who wrote a hypothesis while knowing the label left cues that let a model recover the label from the hypothesis alone, for about 67 percent of SNLI. A rewrite produced under an instruction that names the label is text written with knowledge of the label.

The account cuts both ways. Where the paraphraser misreads the original, it writes the wrong polarity plainly, the classifier follows the text, and the inherited label marks the classifier wrong. Sarcasm is where that should happen most. Sarcasm interpretation, the task of restating a sarcastic utterance literally, was defined by

Peled and Reichart (2017), and it is exactly what a polarity-preserving paraphraser must do on a sarcastic review. SarcasmBench found that current language models underperform supervised pretrained-model baselines at detecting sarcasm across six benchmarks, with GPT-4 the strongest of them (Zhang et al., 2024). Veselovsky et al. (2023) found that when asked to produce sarcastic text, a model fell back on surface markers such as opening with "Oh" or "Wow", and that classifiers trained on its synthetic sarcasm did worse on real sarcasm than classifiers trained on real data. A small paraphraser that misses irony on the way in can turn an ironic negative review into a sincere positive one.

The strongest evidence against this account as a complete explanation is again O'Brien et al. (2024). Their rewriting model, Stable Beluga 2, was less accurate than the fine-tuned BERT on the in-distribution sentiment set (86.53 against 90.85 percent for BERT with rewriting) and on the out-of-distribution average (50.63 against 57.17), yet rewriting still helped BERT. A gain does not require a paraphraser more competent than the classifier. Two differences limit the transfer to the anchor: O'Brien et al. did not tell the rewriting model the label or ask it to preserve one, and they averaged over several rewrites together with the original rather than replacing the original. The annotator account predicts that the anchor's gain should concentrate on items where the paraphraser's own zero-shot label is correct and the classifier's original prediction was wrong, and that the label changes in Section 4.3 should concentrate on items where the paraphraser's own reading is wrong.

5.4 Removing the hard items

The fourth account is that the rewrite does not move text toward one classifier's training register but toward the typical, for every reader. Evidence that model output is less varied than human text is extensive. Munoz-Ortiz et al. (2024) compared human news text with output from six models and found that human text varies more in sentence length and vocabulary and expresses stronger negative emotion, such as fear and disgust, and less joy. Reinhart et al. (2024) found systematic grammatical and rhetorical differences between human and model text that are larger for instruction-tuned models than for base models. Padmakumar and He (2023) found that co-writing with an instruction-tuned model, though not with the base model, made different authors' essays more alike. Kirk et al. (2023) found that reinforcement learning from human feedback reduces output diversity relative to supervised fine-tuning. Zhang et al. (2025) trace part of this collapse to a typicality bias in human preference data, a preference for familiar text that alignment then amplifies. Shumailov et al. (2024) showed that models trained recursively on generated data lose the tails of the original distribution first. Li et al. (2023) found that classifiers trained on model-generated data do nearly as well as those trained on real data for objective tasks and much worse for subjective ones, with an average drop of 27.4 points of macro-F1

for BERT trained on zero-shot synthetic data across the six most subjective tasks they tested, and their diversity analysis found the synthetic data least diverse where the task was most subjective.

If a rewrite turns a mixed review into a plainly positive one, or an idiomatic complaint into a standard one, it does not need to resemble tweets to help a tweet classifier; it has removed the features on which every classifier is weakest. This account differs from alignment in a testable way: it predicts the gain should persist for a classifier trained on Yelp itself, as in the Audinet de Pieuchon et al. (2024) design, and should concentrate on items that annotators would rate as mixed or hedged. It differs from the neutral-rate account in that it predicts a gain in polar-only accuracy too.

One published result complicates the premise that model rewrites are always less varied. Cegin et al. (2023) collected paraphrases for intent classification from ChatGPT and from crowdworkers under the same protocol and found the ChatGPT paraphrases more diverse, with models trained on them at least as robust. The homogenisation result is therefore conditional on the task and the prompt, as the neutralisation result is.

5.5 Two smaller terms

Two further terms can contribute and are not accounts of the effect so much as noise around it. RoBERTa inputs are truncated at 512 tokens in the anchor's code; a long review whose decisive sentence falls after the cut is read without it, and a shorter rewrite may bring that sentence inside the window. The anchor does not report how many reviews exceed the limit. And the fallback rule that assigns neutral to items that fail at inference makes any failure an error; a code comment gives the possibility of unusual characters in model-written text as the reason for a cleaning step before RoBERTa, and the anchor does not report how many items, original or rewritten, reached the fallback. Its released predictions show that none did: by the author's count, no RoBERTa prediction carries the fallback's signature of the label neutral with a score of 0.34. The truncation term is probably small, and its size is not known.

6. Why the two literatures point in opposite directions

The decomposition in Section 3 gives a single reason why rewriting has been found to lower classifier accuracy in one literature and raise it in another, and Table 2 sets the evidence side by side.

The adversarial rows show what happens when the label of the new text is checked. Qi et al. (2021) report that their style-transfer attack succeeded on over 90 percent of SST-2 items against all models, and then had three annotators judge whether 200 adversarial examples kept the original sentiment: 49.5 percent did, against 43 percent for syntactically controlled paraphrases and 3 percent for a GAN-based method. Half of what was counted as the classifier being fooled was the classifier reading a sentence whose sentiment had changed. Wang

Table 2. Studies that score a classifier on text transformed by a model or by people, one row per cited result, with whether the label of the transformed text was checked. Values are as reported by each source.

Source	Transformation	Label of new text	Reported result
Shroff (2026)	polarity-preserving LLM rewrite, Yelp	inherited, not checked	RoBERTa + 3.7 to + 5.8; VADER -1.7 to + 3.2
O'Brien et al. (2024)	in-context LLM rewrites, averaged	inherited, not checked	BERT OOD sentiment + 5.12 mean
Audinet de Pieuchon et al. (2024)	plain LLM paraphrase, Amazon	inherited; classifier fit on rewrites	0.885 to 0.891 (Mistral), 0.899 (GPT-4)
Wang et al. (2025)	GPT rephrasing, 50,000 tweets	no labels; VADER scores	positive -0.031, negative -0.040
Mohammad et al. (2016)	Arabic-English translation	independently re-annotated	manual vs automatic translation labels match about 60 percent
Qi et al. (2021)	style-transfer attack, SST-2	200 checked by humans	success above 90 percent; 49.5 percent kept sentiment
Wang et al. (2021)	14 attack methods, GLUE	human validation	about 90 percent invalid or ambiguous
Morris et al. (2020)	synonym-substitution attacks	human survey	success falls over 70 points under tighter constraints
Kaushik et al. (2019)	human counterfactual revision	new label assigned	models trained on originals fail on revisions

et al. (2021) applied 14 attack methods to GLUE tasks and found that about 90 percent of the resulting examples either changed the meaning or left human annotators disagreeing, so their benchmark kept only the remainder. Morris et al. (2020) found that two synonym-substitution attacks often failed to preserve meaning and that 38 percent of their perturbations introduced grammatical errors; when constraints were tightened to the level human raters needed to judge the meaning unchanged, attack success fell by more than 70 points. Jin et al. (2019) report that human judges still assigned their word-substitution examples the original labels, so the problem is not universal; but in the two studies above that had people check the label, invalid, ambiguous or label-changed examples made up about half or more of what had been counted as the classifier failing.

The constructive rows show what the right design looks like. Kaushik et al. (2019) had people revise reviews to flip their sentiment with minimal edits and labelled the revisions with the new sentiment, and Gardner et al. (2020) had dataset authors perturb test items in ways that typically changed the gold label; both relabelled the new text. Models trained on the original data failed on the revisions, and performance on contrast sets fell by up to 25 percent in some cases. Mohammad et al. (2016) went further and annotated source and translation independently, which is what lets them say both that translation pushes sentiment toward neutral and that, on one of their datasets, manual and automatic translations of the same posts agreed on sentiment only about 60 percent of the time. They also observed that some properties of machine-translated text misled human annotators about the source sentiment while leaving an automatic classifier largely unaffected, which is the label-change term with the opposite sign from the adversarial case.

Behavioural testing already has the right form for a claim about the classifier. The invariance tests in CheckList (Ribeiro et al., 2020) state in advance which perturbations a sentiment model's prediction should

survive, so the assumption that the label is unchanged is written down by the test designer rather than delegated to whatever the perturbation happens to do. A model rewrite used as a test condition has no such statement attached.

Read through the decomposition, the literatures do not disagree about classifiers. The adversarial studies select rewrites the classifier fails on, which selects for label change and inflates measured loss. The anchor instructs its rewrites to state one of two polarities, which can resolve items in the label's favour and inflate measured gain. The neutralisation studies use no labels at all and measure the sentiment-score shift directly. Each design reports D, and D carries the label-change term with whatever sign the design gives it.

Figure 1 places the anchor's four values beside the other accuracy changes measured on rewritten text and beside the neutralisation shift. Every accuracy change in panel (a) was measured against inherited labels; the one study in panel (a) that fitted its classifier to the rewritten text found the smallest changes; panel (b) shows the habit that, under the anchor's scoring, would cost accuracy rather than add it.

7. The calibration, abstention and uncertainty results through the same lens

The anchor's remaining results are offered in support of uncertainty-aware abstention, and they are consistent with that recommendation as a practice. What they show about the classifier is narrower than stated, for reasons the decomposition and the scoring rule explain.

7.1 Lower confidence on sarcastic tweets

The first experiment reports that RoBERTa's mean top-class probability was 0.741 on 867 author-labelled sarcastic tweets and 0.770 on 2,601 others, with standard deviations of 0.166 and 0.169, and concludes that the classifier's confidence registers its own uncertainty on ironic content. The difference is statistically clear at that sample size. It is also small:

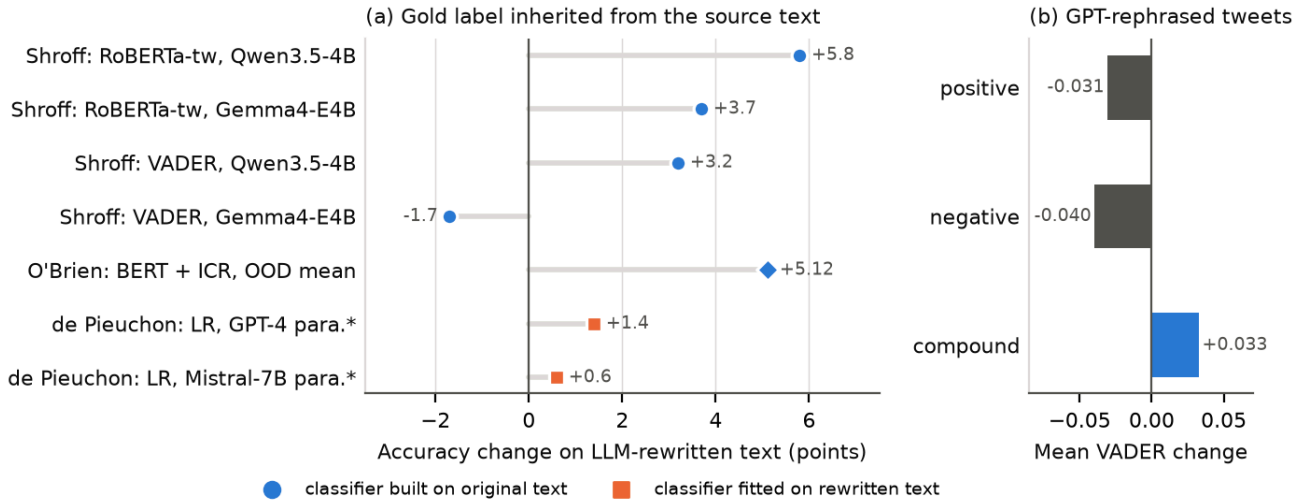


Figure 1. Published effects of model rewriting on sentiment measurement. (a) Change in classifier accuracy, in points, when the test text is rewritten by a language model and scored against the label of the original: Shroff (2026), four classifier and paraphraser pairs, as printed in its Table 2; O'Brien et al. (2024), BERT with in-context rewriting, mean over out-of-distribution sentiment sets; Audinet de Pieuchon et al. (2024), logistic regression fitted and tested on paraphrased Amazon reviews, where the differences from the printed 0.885 are the author's arithmetic (marked *) and lie within the reported intervals. None of these studies re-annotated the rewritten text. (b) Mean change in VADER positive, negative and compound scores after GPT rephrasing of 50,000 tweets, from Wang et al. (2025).

about 0.03 against a spread of about 0.17, roughly a sixth of a standard deviation by the author's arithmetic on the printed values, so the two confidence distributions overlap almost entirely and a threshold on confidence would separate sarcastic from non-sarcastic tweets only slightly better than chance.

More important is what the comparison can license. Lower maximum softmax probability is the generic signature of inputs that are harder or further from the training data, misclassified or out of distribution alike; that is the observation on which the standard baseline for detecting both rests (Hendrycks and Gimpel, 2017). A sarcastic tweet typically pairs positive surface words with a negative situation, and any tweet with mixed surface polarity, ironic or not, should draw a less peaked distribution over positive, negative and neutral. Labelling sarcasm by the author's own statement of intent began with the iSarcasm dataset of Oprea and Magdy (2020), built that way precisely because perceived and intended sarcasm come apart, and iSarcasmEval kept the practice. Oprea and Magdy found that detection models that did well on earlier sarcasm datasets did poorly on their 2020 dataset; that finding concerns their dataset, not the iSarcasmEval set the anchor used, and it bears on the anchor only as evidence about author-labelled sarcasm in general. The tweets labelled sarcastic in iSarcasmEval are therefore not guaranteed to carry the surface cues that would make a classifier hesitate, and the ones that do may hesitate for the surface mixture rather than the irony.

To show that the classifier's confidence tracks its error on ironic content, the experiment would need sentiment labels on the sarcastic tweets, so that confidence could be compared with correctness rather than with a sarcasm flag, and a comparison set of non-sarcastic tweets

matched on surface polarity mixture. The anchor names the first as future work. Without either, the result says that RoBERTa is slightly less confident on tweets their authors call sarcastic. It does not say that the lower confidence is about the sarcasm, or that it falls on the items the classifier gets wrong.

7.2 Calibration on the rewrites

Expected calibration error was 0.056 on the human Yelp text, 0.045 on the Qwen rewrites and 0.029 on the Gemma rewrites. The anchor reads the lower values as following from higher accuracy. ECE compares confidence with accuracy, and accuracy here is computed against inherited labels with neutral predictions always wrong. A neutral prediction made with high confidence on a mixed review is, under this scoring, a confident error, and a rewrite that removes such items lowers ECE whether or not the classifier's probabilities mean anything more. Modern networks are often overconfident, and temperature scaling is the standard post-hoc correction (Guo et al., 2017); Desai and Durrett (2020) found pretrained Transformers calibrated in-domain out of the box but needed a different remedy, label smoothing, for out-of-domain inputs, and Ovadia et al. (2019) found that post-hoc calibration falls short under dataset shift. Those results would lead one to expect ECE to change between originals and rewrites. They do not say which way, and the anchor's numbers cannot separate a better-calibrated reading from a scoring artefact without the neutral rate.

7.3 Abstention and the two uncertainty scores

The abstention result is a selective-classification result, a reject option of the kind Geifman and El-Yaniv (2017) formalised and for which maximum softmax probability is the standard baseline signal (Hendrycks and Gimpel,

2017). Kamath et al. (2020) showed that abstention based on softmax probability alone does poorly under domain shift because models are overconfident on shifted inputs. The anchor's rule was applied to the human-written Yelp text, which is shifted relative to the classifier's training data, and the threshold was set on the evaluation data; the second point in particular suggests that the 6.7-point gain overstates what the same rule would deliver on new data. The flagged items' 41.3 percent accuracy, below binary chance, comes in large part from neutral predictions, which no threshold on confidence could make correct: by the author's count from the released predictions, 269 of the 700 (Section 4.5).

The uncertainty comparison reports that semantic entropy and a sample-disagreement score had nearly identical AUROC, 0.650 and 0.646, for predicting errors on 300 iSarcasmEval tweets, 82 of them sarcastic. Two features of the design make near-identity close to inevitable. Semantic entropy, as introduced by Kuhn et al. (2023) and applied to hallucination detection by Farquhar et al. (2024), clusters free-form generations by meaning before computing entropy; when the generations are single labels from a set of three, the clusters are the labels and semantic entropy reduces to the entropy of the label histogram, as the anchor's own definition states. The second score, which the anchor calls MC-Dropout-style but computes as one minus the share of the modal label among ten temperature samples, is not Monte Carlo dropout in the sense of Gal and Ghahramani (2016), and is a second function of the same histogram. Two summaries of the same ten-sample, three-label histogram, both of which grow as the samples spread, will rank most items identically; this is a property of the histogram, not a finding about either method. And correctness was not scored against a sentiment label, because iSarcasmEval has none. In the released notebook, correctness is agreement between Qwen's modal label and a proxy: "negative" for tweets flagged sarcastic, and RoBERTa's own prediction for the rest; in the released results, 211 of the 300 agree. The AUROC of 0.65 measures how well the spread of the samples predicts disagreement with that proxy, which is mostly disagreement between Qwen and RoBERTa, not accuracy, and the parity between the two methods is a property of the construction rather than evidence that they are interchangeable in general.

8. What the evaluation-confound claim needs

The anchor's most consequential claim is methodological: that a sentiment benchmark augmented, cleaned or paraphrased by language models will systematically overestimate classifier performance relative to organic social-media text, and that generative source should be recorded as metadata. The second half is sound and needs nothing from the first. Recording provenance costs little, and the shares estimated by Liang et al. (2024, 2025) make it overdue. Detecting provenance after the fact is unreliable: Krishna et al.

(2023) showed that paraphrasing machine-written text with a dedicated paraphrase model cut DetectGPT's detection accuracy from 70.3 to 4.6 percent at a fixed false-positive rate, and Yang et al. (2023) showed that rephrased test items pass the n-gram checks used to decontaminate training data. Provenance has to be recorded when the text is made.

The direction is a different matter. "Systematically overestimate" requires that model rewriting move measured accuracy up in general. The evidence in Sections 5 and 6 says the direction depends on at least three things the claim does not fix. It depends on the prompt: a polarity-naming instruction can resolve items in the label's favour, while a plain or even a sentiment-preserving paraphrase prompt, in the Wang et al. (2025) data, pulled sentiment toward neutral. It depends on the paraphraser: within the anchor, VADER gained on one model's rewrites and lost on the other's. And it depends on the scoring rule: a three-class classifier on two-class labels gains from polarisation and loses from neutralisation, while a classifier forced to choose between two labels is affected only through its polar accuracy. A benchmark cleaned by a model told to fix grammar, scored with a binary classifier, could move in either direction. The licensed claim is that model rewriting changes measured accuracy by an amount and in a direction that depend on the prompt, the rewriting model and the scoring rule, and that a benchmark should record all three alongside the provenance flag.

In practice the record would hold five fields per item: whether a model touched the text; which model; the instruction it was given, verbatim; the decoding settings; and, for evaluation sets, whether the item's label was assigned before or after the model touched it. The last field is the one the anchor's design lacks and the one that decides whether a measured shift can be attributed to a classifier. It also separates two cases a single provenance flag merges: text a person wrote and then had a model polish, whose label a person can still vouch for, and text a model rewrote from a labelled original, whose label nobody has checked.

The claim also omits a downstream consequence that is better established than the accuracy shift. Wang et al. (2025) replicated a published regression study of climate-change sentiment on the rephrased tweets and found the same qualitative directions with markedly different effect sizes, including one coefficient that became significant only on the rephrased text. Substantive conclusions drawn from sentiment scores, not only benchmark accuracies, inherit the rewriting term.

9. A licensing procedure

Algorithm 1 states what a study that scores a classifier on model-rewritten text may conclude, given which controls it ran. It encodes which conclusion each published kind of control earns; it proposes no new score.

Algorithm 1. Licensing a claim from a rewrite-shift result. Input: classifier f ; items (x, y) ; rewriter T with prompt P ; reported $D = \text{acc}(f, T(x), y) - \text{acc}(f, x, y)$ Output: the strongest claim the evidence licenses

```

S1: if the labels of  $T(x)$  were never checked:
    claim  $\leftarrow$  AGREEMENT_SHIFT    #  $D$  is a change in agreement with  $y$ 
    if  $P$  names or implies the label: add flag LABEL_IN_PROMPT
S2: if  $f$  has more classes than  $y$ :
    require neutral rate  $n$  and polar accuracy  $a$  on  $x$  and  $T(x)$ 
    if not reported: add caveat NEUTRAL_RATE_UNKNOWN
S3: if a blind re-annotation sample gives  $y_T$  for  $T(x)$ :
    estimate label-change rate  $r = P[y_T \neq y]$ 
    report  $D_{\text{own}}$  on the re-annotated sample
    if  $r$  is small and  $D_{\text{own}}$  is close to  $D$ :
        claim  $\leftarrow$  READING_SHIFT    #  $f$  reads  $T(x)$  better or worse
S4: if the rewriter's own zero-shot label  $z$  is recorded:
    split  $D$  by  $(z == y)$  and by  $(f(x) == y)$ 
    if the gain sits where  $z == y$  and  $f(x) \neq y$ :
        add finding GAIN_FROM_REWRITER
S5: if  $P$  is ablated (label-free, polarity-naming, style-only):
    report  $D$  for each prompt
    if the sign of  $D$  changes: add finding SIGN_SET_BY_PROMPT
S6: if a classifier fitted on the target domain is also scored:
    if  $D$  persists for it: reject ALIGNMENT_WITH_TRAINING_SET
S7: never generalise to all AI text from one  $T$ , one  $P$  and one  $f$ 

```

Steps S1 and S2 cost nothing and apply to every published result in Table 2; under them, the anchor's headline is licensed as a 5.8-point change in agreement with the source label for a three-class classifier on two-class labels, written by a model told the polarity. Step S3 is the one that changes the claim's subject from the label to the classifier, and it is the step no study of model-rewritten text has taken. Step S7 is the one the anchor's discussion does not observe.

10. Measurements that would settle the open part

Each of the controls below can be run on the anchor's released rewrites without generating new ones, except the prompt ablation. Table 3 collects what each account predicts under each control.

Re-annotate a random sample of rewrites blind to the source. A few hundred items, labelled by the same two-class protocol, would estimate the label-change rate and D_{own} . Mohammad et al. (2016) show the design is practical and informative; Qi et al. (2021) show that 200 items are enough to reveal a label-change rate near one half when it exists.

Record the rewriter's own reading. Ask each paraphraser for its zero-shot label on each original, and split the gain by whether that label and the classifier's original prediction were right. If the gain lives where the paraphraser was right and the classifier was wrong, the third account is the main one, and the result is a measurement of the paraphraser.

Ablate the polarity instruction. Rewrite the same reviews with a plain paraphrase prompt, with a prompt that keeps sentiment without naming the two polarities, and with a style-only prompt. Wang et al. (2025) ran a closely related variation on tweets, a plain prompt, a sentiment-keeping prompt and a grammar prompt, and found neutralisation under every one; whether the anchor's gain survives the change is the most direct test of whether the gain belongs to the prompt.

Score a classifier fitted to the reviews' domain. A sentiment classifier fine-tuned on Yelp Polarity training data has nothing to be aligned to. If the gain persists for it, alignment with the Twitter classifier's register is not the explanation; if it vanishes, the alignment account gains its first direct support.

Stratify by item type. Annotate originals as mixed, hedged, ironic or plain, as iSarcasm does for sarcasm with its authors' own labels (Oprea and Magdy, 2020), and report the gain in each stratum. The fourth account predicts it concentrates in the mixed and hedged strata; the annotator account predicts errors concentrated in the ironic one.

11. Objections

"Whatever the mechanism, the practical warning stands: model-touched text changes measured accuracy." Agreed, and Section 8 keeps that part. The objection does not rescue the direction. A practitioner who believes that model-cleaned benchmarks overstate performance will discount high scores and trust low ones; if the cleaning neutralised sentiment, as Wang et al. (2025) found for rephrased tweets, the discount runs the wrong way.

"Label preservation is the paraphraser's job, and the prompt asked for it." The prompt asked for the polarity to be preserved, and a model that complies decides the polarity first. Asking for preservation does not establish that it happened, and the Wang et al. (2025) robustness check is direct evidence that asking does not prevent sentiment from shifting. The anchor's label-change rates of 11.5 to 17.7 percent are changes in the classifier's output, not in the text's label, so they bound neither.

"The Audinet de Pieuchon et al. differences are inside their confidence intervals, so they show nothing." They show nothing about the size of an amplification effect, and the paper says so. They do show that a classifier fitted on rewritten text did not lose accuracy, which is the observation the alignment

Table 3. Predictions of each account under each control. The table is the author's derivation from the accounts as stated in Section 5; "-" means the account makes no definite prediction. The first row's control has been run on the anchor's released predictions (Section 5.2): n fell and polar accuracy rose.

Control	Alignment	Neutral rate	Rewriter as annotator	Hard items removed
Neutral rate and polar accuracy	polar gain	polar flat; n falls	polar gain	polar gain
Blind re-annotation, D_own	D_own near D	D_own near D	lower where z wrong	D_own near D
Split by rewriter label z	uniform	-	gain where z right	-
Prompt without polarity	gain stays	gain shrinks	gain shrinks	gain stays
In-domain classifier	gain gone	gain stays	gain stays	gain stays
Split by mixed or ironic	uniform	on mixed	errors on ironic	on mixed

account needs to be false in order to be the whole story; a gain with no domain gap to close is not alignment, whatever its size.

"O'Brien et al. found large gains with rewriting; that supports the anchor." It supports the existence of rewrite-induced gains, which this paper does not dispute. It undercuts the anchor's mechanism twice: the rewrites did not move toward the in-distribution region, and the rewriting model was less accurate than the classifier yet still helped. Its design, averaging over several rewrites and the original, also differs from replacing the original, so it cannot stand in for the anchor's condition.

"The flip-count inconsistency is a typo." In part it is. Section 4.3 shows from the released predictions that the printed 291 is the net gain and the 282 a remainder, so the count matches the headline once relabelled. The sentence still calls a four-to-one split near-symmetric, and a check whose labels are wrong is a reason to print the transition table beside it.

"Sarcasm is a separate experiment in the anchor; why bring it into the paraphrase result?" Because sarcasm is not confined to tweets. The anchor's own model-generated explanations for flagged Yelp reviews list sarcasm among the ambiguity signals they detect, and one explanation the anchor quotes describes a reviewer mixing sarcasm with genuine complaints. A paraphraser that resolves sarcasm correctly helps the classifier; one that resolves it wrongly writes the wrong polarity in plain words. The anchor measured classifier confidence on sarcastic tweets and paraphrase effects on reviews, and did not measure what the paraphrasers did to sarcastic reviews.

"Data augmentation with model paraphrases is known to help; why would test-time paraphrase be different?" Augmentation helps when added to training data whose labels are known, as Dai et al. (2023) found for few-shot classification with ChatGPT rephrasings and Hartvigsen et al. (2022) for toxicity with model-generated statements. At training time a label-changing paraphrase is noise that more data can absorb. At test time it is the measurement.

12. What is not known

The label-change rate of the anchor's rewrites under human annotation is not known, and no study of model-rewritten review text located for this paper reports one. The zero-shot sentiment accuracy of Qwen3.5-4B and

Gemma4-E4B on Yelp Polarity is not known. How many reviews exceeded RoBERTa's 512-token limit is not known. Whether a classifier trained on Yelp itself gains on the rewrites is not known. Whether the polarisation the anchor's prompt invites and the neutralisation Wang et al. (2025) measured would cancel, add, or depend on the review is not known. Whether the anchor's gain survives a second domain, which the anchor names as future work, is not known.

What is known is narrower and sufficient for the paper's argument: that the anchor's prompt names two polarities, that its scoring makes every neutral prediction wrong, that its flip count is mislabelled, the printed 291 being the net gain rather than the number of improvements, that a fall in neutral predictions carries about half of its largest gain, that its VADER rows change sign with the paraphraser, that the clearest test of the alignment mechanism in the published literature found the rewritten inputs had not moved toward the training domain, and that in the adversarial literature, where the label of rewritten text was checked, about half or more of the rewrites counted as failures had changed it or left annotators unable to agree, though not in every study (Jin et al., 2019).

13. Conclusion

A classifier scored on a model's rewrite of a review is scored against a label that belonged to the review. When the rewrite was written by a model told which of two polarities to keep, the rewrite carries that model's reading of the review, and a gain in agreement with the label can be the rewriter's competence, a fall in neutral predictions that the scoring counted as errors, or the disappearance of the mixed and ironic items that defeat every classifier, as readily as an alignment between the rewrite's style and the classifier's training data. The same structure, with the selection pointed the other way, inflated the fragility the adversarial literature once measured, and its own human checks showed it. The anchor's released predictions show that a fall in neutral predictions carries about half of the 5.8-point gain; the published record does not say which account explains the rest, and it does not license the claim that model-cleaned benchmarks systematically overstate accuracy; a prompt that pulls sentiment toward neutral would push the bias the other way. Three inexpensive measurements, a blind re-annotation sample, the rewriter's own labels, and a prompt ablation, would settle which account holds

for the rest. Until one of them is reported, a rewrite-shift result is a statement about agreement with a label the rewritten text was never given.

References

- Abu Farha, I., Oprea, S. V., Wilson, S., & Magdy, W. (2022). SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 802-814. doi:10.18653/v1/2022.semeval-1.111.
- Audinet de Pieuchon, N., Daoud, A., Jerzak, C. T., Johansson, M., & Johansson, R. (2024). Can Large Language Models (or Humans) Disentangle Text? Sixth Workshop on NLP and Computational Social Science at NAACL 2024. arXiv:2403.16584. doi:10.48550/arXiv.2403.16584.
- Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of EMNLP 2020*. arXiv:2010.12421. doi:10.48550/arXiv.2010.12421.
- Cegin, J., Simko, J., & Brusilovsky, P. (2023). ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. *EMNLP 2023*. arXiv:2305.12947. doi:10.48550/arXiv.2305.12947.
- Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., et al. (2023). AugGPT: Leveraging ChatGPT for Text Data Augmentation. arXiv:2302.13007. doi:10.48550/arXiv.2302.13007. [PREPRINT]
- Desai, S., & Durrett, G. (2020). Calibration of Pre-trained Transformers. arXiv:2003.07892. doi:10.48550/arXiv.2003.07892.
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*. doi:10.1038/s41586-024-07421-0.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ICML 2016*. arXiv:1506.02142. doi:10.48550/arXiv.1506.02142.
- Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., et al. (2020). Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709. doi:10.48550/arXiv.2004.02709.
- Geifman, Y., & El-Yaniv, R. (2017). Selective Classification for Deep Neural Networks. arXiv:1705.08500. doi:10.48550/arXiv.1705.08500.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *ICML 2017*. arXiv:1706.04599. doi:10.48550/arXiv.1706.04599.
- Gururangan, S., Swamyamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. arXiv:1803.02324. doi:10.48550/arXiv.1803.02324.
- Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *ACL 2022*. arXiv:2203.09509. doi:10.48550/arXiv.2203.09509.
- Hendrycks, D., & Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *ICLR 2017*. arXiv:1610.02136. doi:10.48550/arXiv.1610.02136.
- Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). Pretrained Transformers Improve Out-of-Distribution Robustness. arXiv:2004.06100. doi:10.48550/arXiv.2004.06100.
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1). doi:10.1609/icwsm.v8i1.14550.
- Iyyer, M., Wieting, J., Gimpel, K., & Zettlemoyer, L. (2018). Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. *NAACL 2018*. arXiv:1804.06059. doi:10.48550/arXiv.1804.06059.
- Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2019). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv:1907.11932. doi:10.48550/arXiv.1907.11932.
- Kamath, A., Jia, R., & Liang, P. (2020). Selective Question Answering under Domain Shift. arXiv:2006.09462. doi:10.48550/arXiv.2006.09462.
- Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. arXiv:1909.12434. doi:10.48550/arXiv.1909.12434.
- Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2023). Understanding the Effects of RLHF on LLM Generalisation and Diversity. arXiv:2310.06452. doi:10.48550/arXiv.2310.06452. [PREPRINT]
- Kocon, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydlowski, D., Baran, J., et al. (2023). ChatGPT: Jack of all trades, master of none. arXiv:2302.10724. doi:10.48550/arXiv.2302.10724.
- Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. arXiv:2303.13408. doi:10.48550/arXiv.2303.13408.
- Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *ICLR 2023*. arXiv:2302.09664. doi:10.48550/arXiv.2302.09664.
- Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. *EMNLP 2023*. arXiv:2310.07849. doi:10.48550/arXiv.2310.07849.
- Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., et al. (2024). Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. arXiv:2403.07183. doi:10.48550/arXiv.2403.07183.
- Liang, W., Zhang, Y., Codreanu, M., Wang, J., Cao, H., & Zou, J. (2025). The Widespread Adoption of Large Language Model-Assisted Writing Across Society. arXiv:2502.09747. doi:10.48550/arXiv.2502.09747. [PREPRINT]
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. doi:10.48550/arXiv.1907.11692. [PREPRINT]
- Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., & Camacho-Collados, J. (2022). TimeLMs: Diachronic Language Models from Twitter. arXiv:2202.03829. doi:10.48550/arXiv.2202.03829.
- Mohammad, S. M., Salameh, M., & Kiritchenko, S. (2016). How Translation Alters Sentiment. *Journal of Artificial Intelligence Research*, 55, 95-130. doi:10.1613/jair.4787.
- Morris, J. X., Lifland, E., Lanchantin, J., Ji, Y., & Qi, Y. (2020). Reevaluating Adversarial Examples in Natural Language. *Findings of EMNLP 2020*. arXiv:2004.14174. doi:10.48550/arXiv.2004.14174.
- Munoz-Ortiz, A., Gomez-Rodriguez, C., & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57. arXiv:2308.09067. doi:10.48550/arXiv.2308.09067.
- O'Brien, K., Ng, N., Puri, I., Mendez, J., Palangi, H., Kim, Y., Ghassemi, M., & Hartvigsen, T. (2024). Improving Black-box Robustness with In-Context Rewriting. arXiv:2402.08225. doi:10.48550/arXiv.2402.08225.
- Oprea, S., & Magdy, W. (2020). iSarcasm: A Dataset of Intended Sarcasm. arXiv:1911.03123. doi:10.48550/arXiv.1911.03123.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive

- Uncertainty Under Dataset Shift. NeurIPS 2019. arXiv:1906.02530. doi:10.48550/arXiv.1906.02530.
- Padmakumar, V., & He, H. (2023). Does Writing with Language Models Reduce Content Diversity? arXiv:2309.05196. doi:10.48550/arXiv.2309.05196.
- Peled, L., & Reichart, R. (2017). Sarcasm SIGN: Interpreting Sarcasm with Sentiment Based Monolingual Machine Translation. arXiv:1704.06836. doi:10.48550/arXiv.1704.06836.
- Potts, C., Wu, Z., Geiger, A., & Kiela, D. (2021). DynaSent: A Dynamic Benchmark for Sentiment Analysis. arXiv:2012.15349. doi:10.48550/arXiv.2012.15349.
- Pyreddy, S. R., & Zaman, T. S. (2025). EmoXpt: Analyzing Emotional Variances in Human Comments and LLM-Generated Responses. IEEE CCWC 2025. arXiv:2501.06597. doi:10.48550/arXiv.2501.06597.
- Qi, F., Chen, Y., Zhang, X., Li, M., Liu, Z., & Sun, M. (2021). Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. EMNLP 2021. arXiv:2110.07139. doi:10.48550/arXiv.2110.07139.
- Reinhart, A., Markey, B., Laudendach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2024). Do LLMs write like humans? Variation in grammatical and rhetorical styles. arXiv:2410.16107. doi:10.48550/arXiv.2410.16107. [PREPRINT]
- Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP models with CheckList. arXiv:2005.04118. doi:10.48550/arXiv.2005.04118.
- Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. arXiv:1912.00741. doi:10.48550/arXiv.1912.00741.
- Shroff, S. (2026). When AI Rewrites, Classifiers Relax: Uncertainty-Aware Sentiment Analysis on Sarcastic and AI-Paraphrased Social Text. arXiv:2608.15338. doi:10.48550/arXiv.2608.15338. [PREPRINT]
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature. doi:10.1038/s41586-024-07566-y.
- Veselovsky, V., Ribeiro, M. H., Arora, A., Josifoski, M., Anderson, A., & West, R. (2023). Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. arXiv:2305.15041. doi:10.48550/arXiv.2305.15041. [PREPRINT]
- Wahle, J. P., Ruas, T., Kirstein, F., & Gipp, B. (2022). How Large Language Models are Transforming Machine-Paraphrased Plagiarism. arXiv:2210.03568. doi:10.48550/arXiv.2210.03568.
- Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A. H., & Li, B. (2021). Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. NeurIPS 2021 Datasets and Benchmarks. arXiv:2111.02840. doi:10.48550/arXiv.2111.02840.
- Wang, Y., Eshghi, A., Ding, Y., & Gopal, R. (2025). Echoes of authenticity: Reclaiming human sentiment in the large language model era. PNAS Nexus, 4(2). doi:10.1093/pnasnexus/pgaf034.
- Wang, Z., Xie, Q., Feng, Y., Ding, Z., Yang, Z., & Xia, R. (2023). Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. arXiv:2304.04339. doi:10.48550/arXiv.2304.04339. [PREPRINT]
- Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., & Stoica, I. (2023). Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. arXiv:2311.04850. doi:10.48550/arXiv.2311.04850. [PREPRINT]
- Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., & Sun, M. (2023). Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLMs Evaluations. NeurIPS 2023 Datasets and Benchmarks. arXiv:2306.04618. doi:10.48550/arXiv.2306.04618.
- Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., & Shi, W. (2025). Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. arXiv:2510.01171. doi:10.48550/arXiv.2510.01171. [PREPRINT]
- Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005. doi:10.48550/arXiv.2305.15005. [PREPRINT]
- Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. NeurIPS 2015. arXiv:1509.01626. doi:10.48550/arXiv.1509.01626.
- Zhang, Y., Zou, C., Lian, Z., Tiwari, P., & Qin, J. (2024). SarcasmBench: Towards Evaluating Large Language Models on Sarcasm Understanding. arXiv:2408.11319. doi:10.48550/arXiv.2408.11319. [PREPRINT]