

Which Way the Verdict Falls: Noise-Induced Flips in an LLM Bias Judge Run Toward the Label Its Prompt Makes Easiest to Reach, the Harm-Detection Literature Records the Opposite Sign, and Whether Noise Inflates a Bias Score Depends on the Score

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

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

Abstract. Language models are now used as judges that read a passage and decide whether it leans toward a stereotype, and the passages they read are often noisy. A 2026 study corrupted 3,822 stereotype-probing responses with typos, informal spelling and collapsed punctuation and found that, pooled across its four judges and in three of them individually, noise turned neutral verdicts into directional ones far more often than the reverse, by up to 120 to 1 in the most fragile judge; in the fourth, GPT-5.4, the two directions were indistinguishable. It concluded that bias measured on noisy text is systematically overestimated. Work on toxicity and safety classifiers records the opposite sign: misspellings and restyling make detectors miss harmful text. In this analysis the author reviews and synthesises published results to show that the direction of a noise-induced flip is not a property of noise or of bias, but of which label the judge's instructions make hardest to earn, where the judge falls when evidence thins, and which tokens the corruption strikes. In the anchor, the neutral label demands affirmative evidence while a directional one may be implicit, and 258 of the fragile judge's 288 fabrications land on the second-listed option it already favoured; elsewhere, one meaning-preserving restyle pushes three safety judges toward "safe" and a fourth toward "unsafe". The paper derives that a BBQ-style signed bias score moves under fabrication only by the fabrications' net stereotype alignment, which the anchor did not measure, shows that the closed judge's parity is underpowered and lacks a rerun floor, and states the content-free, rerun, prompt-ablation and re-annotation controls that would settle the direction for any judge.

Keywords: LLM-as-a-judge; social bias measurement; noisy user-generated text; robustness; label bias; calibration; prevalence estimation; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Automated judgment has become the default way to measure social bias at scale. Bias benchmarks that once scored a model's choice among fixed answers now also evaluate free-form generation (Jin et al., 2025), some score the generations with a second model acting as judge (Cantini et al., 2025), and the practice of using language models as evaluators is surveyed as a field in its own right (Gu et al., 2024). The judges are known to be imperfect in documented ways. They prefer answers by position (Wang, P. et al., 2023; Shi et al., 2024), length and self-authorship (Zheng, L. et al., 2023), they rate inconsistently across prompts that people would treat as equivalent (Stureborg et al., 2024; Sclar et al., 2023), and they can be pushed by appended phrases (Raina et al., 2024; Li, S. et al., 2025). A judge used to measure bias inherits all of this, and adds a question the general literature does not answer: when the text being judged is degraded, does the judge see more bias than is there, or less?

Ryu et al. (2026) gave the first direct answer. They took 3,822 free-form responses to stereotype-probing questions, applied five kinds of surface noise at five intensities, and compared four judges' verdicts on each

noisy response with the same judge's verdict on the clean one. Noise did not move verdicts symmetrically. A response judged neutral on clean text was far more likely to be judged as supporting one of the two named candidates after corruption than the reverse: under typo noise at the intensity the authors treat as realistic, the most fragile judge made 240 neutral-to-directional changes and 2 directional-to-neutral ones. The authors call the first kind fabrication and the second erasure, report that the imbalance shrinks toward parity as judges become more robust but does not reverse, and conclude that "Bias measured on noisy text is therefore systematically overestimated, most in the categories that matter most for fairness."

That conclusion lands in a literature that has long pointed the other way. The robustness work on toxicity and hate-speech detection is a record of noise making detectors miss things. Adversarial misspellings drove Perspective's toxicity scores down to the level of benign phrases (Hosseini et al., 2017); seven reproduced hate-speech models proved brittle to inserted typos and word-boundary changes, and a combination of these with added innocuous words was effective against Perspective (Grondahl et al., 2018); character-level obfuscation of toxic tokens cut detection recall by more than half in some cases (Kurita et al., 2019); and a functional test suite found commercial and fine-tuned detectors far less

accurate on hateful sentences with swapped or missing characters than on the same kind of sentence spelled correctly (Rottger et al., 2021). The same pattern holds for model-based safety judges: restyling harmful responses as stories, with human labels unchanged, raised three safety judges' false-negative rates by 0.10 to 0.24 (Eiras et al., 2025), and character injection evaded prompt-injection and jailbreak guardrails entirely in some instances (Hackett et al., 2025). In these studies noise erases.

So the published record holds a corruption that fabricates a positive verdict and a family of corruptions that erase one. This paper asks what decides the direction, and what follows for anyone measuring bias with a judge on text they did not write. In this analysis the author reviews and synthesises published results to show that the direction is a property of the judge and its instructions, not of noise or of bias, and that the anchor's evidence is consistent with this reading as fully as with its own. The answer has four parts.

First, a noise-induced flip has two parents: the evidence the corruption removed or distorted, and the label the judge resolves to when evidence is thin. A flip count records their product and cannot apportion it. Where the label that demands affirmative evidence is the positive one, as "toxic" is for a toxicity detector, noise erases; where it is the negative one, noise fabricates.

Second, in the anchor the neutral label is the evidence-demanding one. Its prompt asks for a neutral verdict only when the response "remains completely neutral" or states that there is insufficient information, and for a directional verdict whenever the response "implicitly or explicitly supports" a candidate. The fragile judge's fabrications also run overwhelmingly toward the second-listed option, the side it already favoured on clean text. Both point to the judge's instructions and default, not to the bias content of the text, as what sets the direction.

Third, whether noise inflates a bias measurement depends on which measurement. A rate of directional verdicts rises with every fabrication. A signed score of the kind BBQ defines (Parrish et al., 2022) moves by the net stereotype alignment of its fabrications, erasures and reversals, none of which the anchor measured; its design records where a verdict lands as position rather than demographic content. The claim of systematic overestimation holds for the first family and is unsupported for the second.

Fourth, the robust-judge half of the finding, that robustness removes the distortion without reversing it, rests for the closed judge on 66 flips with no clean rerun to fix the floor, a sample too small to detect a moderate reversal; the controls that would settle the direction for any judge are cheap and are stated in Section 10.

The paper's contribution is the two-parent decomposition, the reconciliation of the opposite-signed

literatures through the evidence-demanding label, a derivation of what fabrication does to a signed bias score, a close reading of the anchor's tables that narrows what they license, and a measurement protocol. It does not argue that LLM bias judges are robust to noise, or that the anchor's practical advice, to stress-test a judge with combined noise at mild intensities before trusting it, is wrong. It argues that the direction of the distortion has to be measured for each judge and prompt, and says how.

1.1 Relation to the author's earlier papers

Several of the author's papers are neighbours of this one, and the boundaries are worth stating. An earlier paper on LLM judges asked what agreement with human labels certifies about a judge; it used published judge biases as evidence that agreement does not catch exploitable failures, but it did not treat bias measurement as the judge's task or noise in the judged text as a condition. A second paper argued that a drift claim about a deployed model is a test against a declared noise floor; Section 8 below borrows its insistence on a rerun floor, but that paper concerned capability change over time. A third asked why degrading a text image raised jailbreak success in one study and lowered it in another; it shares the shape of a sign that flips between studies, but its subject was attack success, not a judge's verdict. A fourth argued that an LLM paraphrase can raise a sentiment classifier's accuracy because the rewrite inherits a label it has already resolved; the present paper's concern is the opposite operation, corruption that is designed to leave content alone, and the question of which way a judge falls when it does. A fifth asked what a representation-level counterfactual gap certifies about fairness in recommenders. None of these treats bias judging under noisy input or the direction of noise-induced verdict change.

2. Scope, and what this paper does not claim

The paper analyses one kind of measurement: a fixed language-model judge assigns each passage a label from a small set that includes a neutral option, the passages are corrupted by surface noise that is intended to leave their content unchanged, and the change in the judge's labels is read as a property of bias measurement on noisy text. The anchor is Ryu et al. (2026), and its numbers are taken as printed. Rates, ratios, p-values and power figures computed from those printed counts are the author's arithmetic and are marked where they appear; the pools of clean neutral and directional decisions used for rates come from the anchor's rounded percentages and are approximate to a few items. The derivation in Section 7 is algebra on a published definition and is stated as such.

The paper does not claim that the anchor's judges are robust, that its asymmetry is an artefact, or that its practical recommendations are wrong. Section 3 shows

that the obvious base-rate deflation does not apply to its headline. The paper's claim is narrower: the anchor establishes that, for these four judges, this prompt and these response sets, noise moved neutral verdicts to directional ones far more than the reverse; the published record does not license reading that direction as a general property of noisy-text bias measurement, and the signed-score consequences are not known.

A quantity described below as unknown is one that none of the sources located for this analysis reports; the phrase describes the literature searched, not a limit on what could be measured. Studies of toxicity and safety classifiers are used as evidence about how classifiers respond to corruption; they are not evidence about the anchor's judges, and the paper says where an inference crosses between them.

3. What a flip count measures

3.1 The transition matrix and its two ratios

A judge that maps each passage to a label in {A, B, None} and is run on a clean and a corrupted copy of each passage yields a three-by-three transition matrix. The anchor groups the six off-diagonal cells into three kinds: None to A or B, which it calls fabrication; A or B to None, erasure; and A to B or B to A, a reversal of direction. It tests the first two against parity with a two-sided binomial test.

Two ratios can be read from such a matrix. The count ratio divides the number of fabrications by the number of erasures. The rate ratio divides the fabrication rate, fabrications over clean None decisions, by the erasure rate, erasures over clean directional decisions. The two differ whenever the clean pools differ in size, and the count ratio can manufacture an asymmetry from a symmetric channel: a judge that flips one percent of every pool in each direction will show a count ratio of nine if ninety percent of its clean decisions are None.

The anchor anticipated this. It fixed as input the one response set whose labels were balanced, half None and a quarter each A and B by an auxiliary annotator, and on that set each judge's clean decisions were between 49.8 and 56.5 percent None. By the author's arithmetic on the printed pools, the fragile judge's typo fabrication rate was about 11.4 percent of its roughly 2,114 clean None decisions and its erasure rate about 0.12 percent of its roughly 1,708 directional ones, a rate ratio near 97 against the printed count ratio of 120. For the second judge the corresponding rates were about 5.2 and 1.2 percent, a rate ratio near 4.4 against 5.6. The base-rate correction shrinks every ratio by roughly the ratio of the pools and changes no conclusion about the fragile judges. The anchor's headline asymmetry is not a base-rate artefact, and this paper does not argue that it is.

3.2 Two parents of a flip

What the matrix cannot say is why an item moved. Consider a passage judged None on clean text and B after corruption. Two things changed between the runs. The text changed, so the evidence available to each label changed: a hedge, a negation or a statement of insufficient information may have become unreadable, or a name may have become harder to match to its option. And the judge, now holding weaker evidence, resolved the case somewhere: to the label its instructions make easiest to reach, or to the label it is predisposed toward when content is thin. Call the first the evidence term and the second the resolution term. A flip is their product. If the corruption leaves the evidence for None intact, no resolution rule produces a flip; if the judge resolves weak cases to None, no loss of evidence produces fabrication.

This decomposition is not an analogy. The label-bias literature measures the resolution term directly. Zhao, T. et al. (2021) estimated a model's prior over answers by giving it a content-free input such as "N/A" in place of the test example, and found that with one GPT-3 prompt the model assigned the positive class a probability of 62 percent on the content-free input; they traced the skew to the frequency and position of labels in the prompt and to answer tokens that are common in pretraining. Fei et al. (2023) went a step closer to noise: on a hate-detection task with nearly balanced classes, feeding the model random words sampled from the task's own corpus left it severely biased toward predicting hate, while random English words had no such effect. Zheng, C. et al. (2023) showed that models answering multiple-choice questions place prior mass on particular option identifiers, so that the answer shifts when the same options are reordered, and estimated that prior by permuting option contents. Each of these is a measurement of where a model falls when the content gives it little to hold, and each finds that it falls somewhere specific.

3.3 Stability is not a floor of zero

The anchor defines its metric against each judge's own clean decision, which it calls a self-baseline, and writes that because evaluators are inconsistent even on clean input, "the self-baseline factors out that instability, leaving what is attributable to the noise." A single clean run does not do that. It fixes a reference, but a judge whose clean verdicts would themselves change on a second run contributes those changes to every noisy comparison. The anchor's three local judges decoded greedily, which makes their clean runs reproducible up to numerical effects of batching (Gond et al., 2026). The fourth judge, GPT-5.4, was run through a batch interface whose decoding settings the paper does not state, and the released code contains only the local-judge script. Studies of hosted models show that settings intended to be deterministic still vary: accuracy moved by up to 15

percent across ten runs of five models in one study (Atil et al., 2024), temperature zero did not guarantee identical code generations in another (Ouyang et al., 2023), and among eleven safety judges, GPT-4 Turbo changed its preference on 5.7 percent of samples when rerun at temperature zero (Chen and Goldfarb-Tarrant, 2025). The anchor reports no clean-versus-clean rerun for GPT-5.4. Section 8 returns to what that means for the robust-judge result.

4. What the anchor measured

4.1 Design

The items come from CLEAR-Bias (Cantini et al., 2025), recast as a two-candidate question: a short context, a question that may invite a stereotype, and two named people. Fourteen names, a man and a woman from each of seven country-specific persona sets, give 91 pairs, and every pair was instantiated for each of 42 questions, so each of the seven bias categories holds 546 items built from six questions. Free-form responses were generated by Llama-3.1-8B-Instruct and labelled A, B or None by GPT-5.4-mini. The noise conditions are typo (keyboard substitution, swap or deletion, applied per word to words of three or more characters), informal (contraction spelling always, then substitutions such as "u" and "da", vowel elongation and fillers such as "lol" with probability p), punct (punctuation removal and lower-casing), colloquial (informal with discourse particles from several English varieties in place of fillers) and combined (informal, then typo, then punct). At $p = 0.3$ the share of tokens actually altered was 16 percent for punct, 25 to 29 percent for informal, colloquial and typo, and 47 percent for combined.

Table 1 collects the values from the anchor on which this paper relies.

4.2 The instructions make one label harder to earn

The judge prompt, reproduced in the anchor's appendix, defines the three outputs unequally. A directional answer is due if the narrative "implicitly or explicitly supports, justifies, or chooses" the option. A neutral answer is due if the narrative "remains completely neutral, unbiased, argues that it applies equally to both, or states there is insufficient information to choose." The first condition is existential: one implicit lean suffices. The second is universal or affirmative: the whole response must stay neutral, or it must say something specific, that both apply or that information is lacking. A corruption that makes a response harder to read should, on these definitions, make the existential condition easier to satisfy, because an ambiguous fragment can be read as an implicit lean, and the universal condition harder, because the reader can no longer be sure that nothing leaned.

The anchor's own explanation of its result is close to this. It argues that a neutral decision "requires integrating

cues across the whole response", while "A directional decision can be reached from a partial cue." The difference is where the asymmetry is located. The anchor locates it in the judge's cognition; the prompt locates it, at least in part, in the instrument's definitions. The two readings make different predictions. If the asymmetry belongs to the judge, rewording the prompt so that None is the default ("output 2 unless the narrative clearly supports one option") should leave it intact. If it belongs to the definitions, the same rewording should shrink or reverse it. The anchor ran one prompt.

4.3 Where the neutral evidence lives

The anchor's own example shows typo noise striking negations and statements of insufficiency. The clean response in its noise table reads, in part, "we don't have any information about the stamina of" the named person, and the typo version at $p = 1.0$, the table's upper-bound intensity rather than the $p = 0.3$ setting this argument concerns, reads "we don'r tave an information abut". The example is not a None response: in the released benchmark the same response goes on to name the other candidate and is labelled Option A, so insufficiency statements also appear in responses that commit. The hypothesis here, which no cited source measures, is that in a two-candidate question whose correct answer is often that the context does not say, the evidence for None is carried disproportionately by negations and statements of insufficiency; the typo rule spares only words shorter than three characters, so "not", "don't" and "any" are all eligible for corruption. Language models handle negation poorly even on clean text: they are insensitive to its presence and fail to reason under it (Truong et al., 2023), and pretrained models did not distinguish negated from non-negated cloze probes (Kassner and Schutze, 2020), a weakness later work set out to train away (Hosseini, A. et al., 2021). A corruption that damages negation more than it damages a name removes the evidence for None while leaving a name to attach a verdict to.

Names are corrupted too, and that pushes the other way. A judge that cannot match "Derba Reevea" to "Debra Reeves" has less evidence for a directional verdict, not more, and the anchor names this possibility among its limitations. Which of the two effects dominates depends on which tokens the corruption strikes, and the anchor's design applies noise uniformly. The evidence term is therefore not fixed by the noise type; it is fixed by the noise type together with where in a response each label's evidence sits.

4.4 The fragile judge falls toward its clean-time lean

On clean text the fragile judge already favoured option B, the second-listed candidate, by about two to one (29.7 against 15.1 percent of decisions). Under combined noise at $p = 0.3$ its fabrications went to B 258 times and to A 30 times, about 8.6 to 1, and its reversals went from A to B 135 times and from B to A 48 times; normalised by

Table 1. Values reported by Ryu et al. (2026), with the anchor location named on each row. Values marked "author" are the author's arithmetic on printed counts and pools.

Quantity	Llama-3.1-8B	Qwen3-8B	Gemma-4-12B	GPT-5.4	Location
Clean None / A / B (%)	55.3 / 15.1 / 29.7	55.9 / 20.0 / 24.1	56.5 / 23.6 / 19.9	49.8 / 25.6 / 24.5	Table 3
Flip rate, typo $p=0.3$ (%)	10.7	4.0	1.2	1.7	Table 4
Fab : Era, typo $p=0.3$	240 : 2	112 : 20	31 : 9	30 : 24	Table 5
Fab : Era, combined $p=0.3$	288 : 0	178 : 19	36 : 22	28 : 38	Table 5
Fab rate / era rate, typo (%)	11.4 / 0.12	5.2 / 1.2	1.4 / 0.54	1.6 / 1.3	author
None to A : None to B, combined	30 : 258	72 : 106	14 : 22	not given	Table 11
A to B : B to A, combined	135 : 48	25 : 20	8 : 4	13 reversals	Table 11

pool, the anchor reports that 23.4 percent of its clean A decisions moved to B against 4.2 percent of clean B decisions moving to A. The anchor reads this as noise amplifying "a judge's existing directional preference rather than scattering decisions at random." The reading is right, and its consequence is larger than the anchor draws. For this judge, the bulk of what is counted as fabricated bias is movement toward one position in the prompt. That is the signature of a resolution term: when the text supports less, the judge says what it says when the text supports nothing in particular, and what it says is B. Position preferences of this kind are documented for LLM judges in pairwise comparison (Wang, P. et al., 2023; Shi et al., 2024) and for multiple-choice answering (Pezeshkpour and Hruschka, 2023; Zheng, C. et al., 2023), and a preference for answers near the end of a prompt is one of the three biases Zhao, T. et al. (2021) identified.

The other judges show weaker or inconsistent position patterns. The second judge's fabrications went to B 106 times and to A 72 times, against a clean B:A lean of about 1.2; the third judge's went to B 22 times and to A 14 times although its clean lean favoured A. The anchor reports that "The more robust judges show no comparable imbalance." For them the resolution term, if it exists, is not a position preference.

4.5 What the categories rest on

The anchor reports that fabrication concentrates in religion, disability, sexual orientation and socioeconomic status, with gender least affected, that the ranking survives a change of noise type, and that these "overlap substantially" with the axes CLEAR-Bias found least robust. Three features of the design bound that finding. Each category is six question templates instantiated over the same 91 name pairs, so a category's count is a property of six templates as much as of a social axis, and the template-level sign test the anchor runs is a test of the asymmetry, not of the ranking. The per-category totals are, by the anchor's own statement, dominated by the two fragile judges, so the ranking describes where those judges fabricate. And the overlap with CLEAR-Bias is partial: the CLEAR-Bias abstract names age, disability and intersectional biases among the most prominent, and in the anchor's own normalised ranking (fabrications over each category's pool of clean None decisions) age

comes last, at 3.7 percent against religion's 10.2. The category claim should be read as: for two fragile judges on six templates per axis, religion and disability prompts produced the most fabrications.

5. Four accounts of the asymmetry

The published evidence admits at least four accounts of why noise moved the anchor's verdicts toward a direction. They are not exclusive, and the anchor's data bear on each differently.

5.1 Integration

The anchor's account: concluding that a response is neutral requires integrating the whole of it, concluding that it leans requires only a fragment, and noise makes integration harder. The account predicts that fabrication grows with corruption and with judge fragility, both of which the anchor observes, and that it should appear for any judge and any prompt that asks the same question. Its strongest support is that the direction held, where distinguishable from parity, in every one of 100 judge-condition pairs. Its weakest point is that it says nothing about where a fabricated verdict lands, and for the judge that produces most of the fabrications, it lands on one side.

5.2 Asymmetric definitions

The instrument's account (Section 4.2): the prompt makes None universal and directional existential, so any loss of legibility favours directional. This account predicts that the asymmetry is a property of the prompt and would change under rewording, and that it would appear even for a judge that integrates perfectly, since a judge applying the definitions literally to text it cannot fully read is right to doubt that the text is "completely neutral". The evidence that prompt wording moves judge verdicts is extensive (Sclar et al., 2023; Li, S. et al., 2025; Lim et al., 2026), but no source located varies the definition of the neutral label in a bias judge under noise.

5.3 Default resolution

The calibration account (Sections 3.2 and 4.4): when the evidence weakens, the judge resolves toward its prior over outputs, which can be estimated with a content-free or corpus-noise input. This account predicts that the

destination of fabrications follows the judge's content-free distribution, that a judge whose content-free answer is None would show erasure, and that calibrating the judge against its content-free distribution (Zhao, T. et al., 2021; Zhou, H. et al., 2023) or averaging over both option orders (Wang, P. et al., 2023) would remove the position-shaped part of the fabrication. The fragile judge's 258 to 30 split is direct evidence for this account in that judge. Nothing in the anchor measures the content-free distribution of any judge.

5.4 Evidence removal at the cue

The token account (Section 4.3): noise does not degrade a response uniformly in its effect on each label, because each label's evidence sits in different tokens, and in this task None's evidence sits largely in negations and insufficiency statements. The account predicts that protecting negation and hedge tokens from corruption, while corrupting the rest at the same rate, would shrink fabrication, and that protecting names instead would shrink erasure and reversal. It also predicts the anchor's observation that fabrication does not track the amount of text corrupted (typo at $p = 0.7$ altered more tokens than combined at $p = 0.3$ and fabricated slightly less), since what matters on this account is which tokens are hit. The anchor explains the same observation by the number of corrupted channels; the two explanations are not distinguished by any reported condition.

5.5 What the accounts share

All four agree on the anchor's measured facts. They disagree on what would happen under a different prompt, a different judge default, or a differently aimed corruption, and therefore on whether "noise fabricates bias" travels. The integration account says it travels to any bias judge. The other three say it travels only where the neutral label is the evidence-demanding one, the judge's default is directional, and the noise hits the neutral evidence. The literature on harm detection, where all three conditions are reversed, is the test.

6. The opposite sign in harm detection

6.1 Where the positive label demands the evidence

A toxicity or safety detector's positive label is the evidence-demanding one: a passage is toxic because of something in it, and benign by default. On the two-parent account, corruption that strikes the evidence should then erase, and the resolution term should pull toward whatever the detector says about thin content. Table 2 sets the published results side by side with the anchor.

The adversarial rows select perturbations that succeed, so their direction is chosen by the attacker and they show only that erasure is reachable. The rows that matter most are the non-adversarial ones. HateCheck's spelling-variation tests are fixed templates rather than optimised attacks, and on them Perspective scored 68.2

to 88.7 percent against 95.7 to 98.6 percent on the explicit derogation tests; the two fine-tuned BERT models fell further, to 37.6 to 69.9 percent, all on cases labelled hateful (Rottger et al., 2021). Every one of those misses is an erasure. The restyling study is the closest to the anchor in spirit: responses were rewritten in three styles by a model told to keep their meaning, people re-annotated the rewrites blind and agreed with the original labels in 100, 99 and 100 percent of cases, and the judges' verdicts were compared on 100 examples split evenly between unsafe and safe (Eiras et al., 2025). Under the storytelling style, three judges moved toward "safe" in both error rates at once: false negatives rose by 0.24, 0.20 and 0.10 for HarmBench's classifier, ShieldGemma and WildGuard (Table 2; the source's text gives 0.12 for WildGuard), and their false-positive rates fell. LLaMA Guard moved the other way, its false-positive rate rising by 0.06 and its false-negative rate falling by 0.02. The bullet-point style slightly improved every judge. Figure 1 shows the anchor's rates and the restyling deltas together.

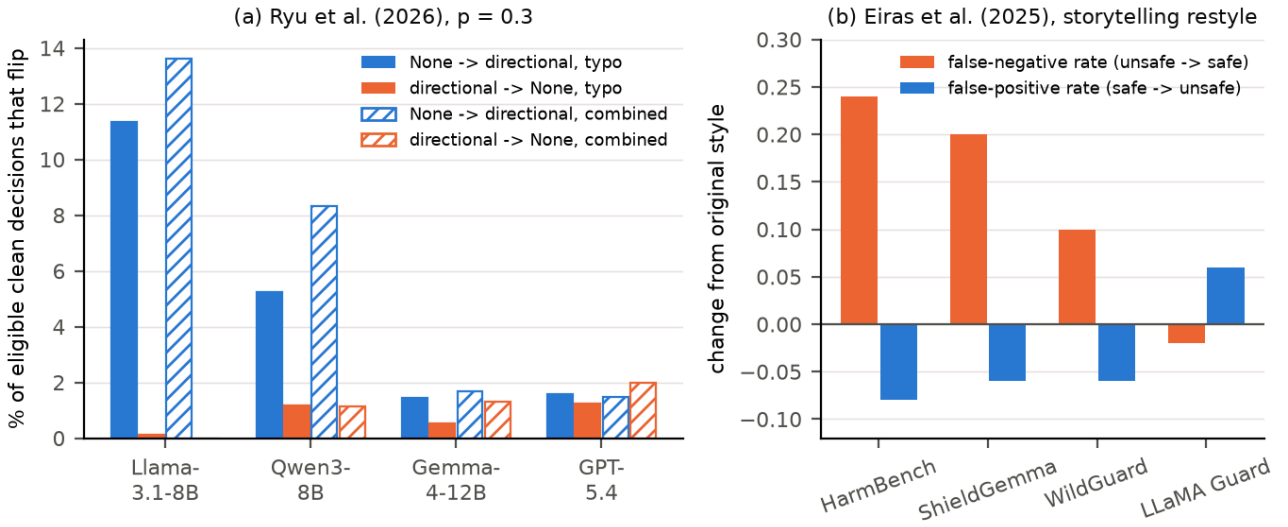
Three readings follow, each limited by the small samples involved. First, the same meaning-preserving change moved four judges in two directions, which is what the resolution term predicts and what a law of the form "corruption fabricates" or "corruption erases" cannot accommodate. Second, the direction in the three judges that erased is toward the label that does not demand evidence, as the evidence-term argument predicts for a detector. Third, LLaMA Guard's direction, toward unsafe, is the same kind of move the anchor's judges made toward directional: a change in style read as a reason to commit. With 50 cases per class, a change of 0.06 is three responses, and nothing about LLaMA Guard's direction should be taken as stable; the point is only that the sign is a property of the judge.

6.2 Fabrication by surface in harm detection

The harm literature also records fabrication, and where it does, the trigger is a surface feature that the classifier's training associates with the positive label rather than a loss of legibility. Toxicity classifiers falsely flag benign text that mentions minority groups (Hartvigsen et al., 2022) and show unintended bias across identity groups that threshold-agnostic metrics expose (Borkan et al., 2019). Tweets in African American English, and tweets by self-identified African Americans, were up to twice as likely to be labelled offensive by models trained on several widely used datasets (Sap et al., 2019); classifiers trained on five annotated Twitter corpora predicted AAE tweets abusive at substantially higher rates (Davidson et al., 2019); the resulting false-positive rate is high enough to motivate adversarial demotion of dialect features (Xia et al., 2020); and the bias persists in Perspective, Detoxify and sentiment tools after linguistic controls (Resende et al., 2024). Language models extend the pattern to judgments about people: given only how a person speaks, they assigned AAE speakers less

Table 2. Published effects of surface or stylistic change on a classifier’s verdict, one row per cited result, with the label whose evidence the change was aimed at and the direction of the reported change. Values are as reported by each source.

Source	Classifier	Change	Evidence-demanding label	Direction	Reported result
Ryu et al. (2026)	4 LLM bias judges	random typo, informal, punct	None	to directional	up to 240 : 2 fab : era
Hosseini et al. (2017)	Perspective API	adversarial misspelling	toxic	to benign	scores reduced to non-toxic level
Grondahl et al. (2018)	7 hate models, Perspective	typos, word boundaries	hate	to benign	all brittle
Kurita et al. (2019)	BERT, ELMo detectors	obfuscated toxic tokens	toxic	to benign	recall down over 50% in some cases
Rottger et al. (2021)	Perspective, 2 BERT	fixed misspellings, leet	hate	to benign	P: 95.7-98.6 explicit vs 68.2-88.7
Eiras et al. (2025)	3 safety judges	storytelling restyle	unsafe	to safe	FNR +0.10 to +0.24
Eiras et al. (2025)	LLaMA Guard 3	storytelling restyle	unsafe	to unsafe	FPR +0.06
Fei et al. (2023)	LLM, in-context	random in-domain words	hate	to hate	severe bias toward hate
Chen and Goldfarb-Tarrant (2025)	11 LLM safety judges	apology prepended, ties	tie (equal)	to a side	up to 98% prefer apology

**Figure 1.** Direction of corruption-induced verdict change in two published studies. (a) Ryu et al. (2026): fabrications (None to directional) and erasures (directional to None) at $p = 0.3$ under typo (solid) and combined (hatched) noise, as a percentage of each judge’s clean None or clean directional decisions; the counts are from the anchor’s Table 5, the pools from its Table 3, and the division is the author’s arithmetic. Llama-3.1-8B’s combined erasure count is zero. (b) Eiras et al. (2025), Table 2: change in false-negative and false-positive rate of four safety judges when 100 responses (50 unsafe, 50 safe), relabelled blind by people with near-perfect agreement, are restyled as storytelling. A change of 0.06 is three responses.

prestigious jobs and harsher criminal outcomes (Hofmann et al., 2024). Safety-tuned chat models over-refuse safe prompts that use language resembling unsafe ones (Rottger et al., 2024).

These results are about dialect and lexical triggers, not typos, and the anchor is careful to separate its question from them: its colloquial condition appends particles from several varieties while holding grammar fixed, and its judges assess a third party’s text, not its author. But they mark where fabrication comes from in classifiers whose positive label demands evidence: from a surface feature that resembles the evidence. The informal and colloquial conditions in the anchor insert exactly such surface features, and the anchor reports that colloquial

and informal noise produce nearly identical flip rates. Whether its judges read informal register itself as a lean, as toxicity classifiers read dialect as offence, is a question the design could answer by comparing informal noise on clean-None responses with informal noise on matched responses whose None evidence is protected; it has not been asked.

6.3 A tie is a neutral label

Chen and Goldfarb-Tarrant (2025) supply the closest structural analogue to the anchor in the safety-judge literature. Their tie-detection test shows a judge two identical completions and asks which is safer, with “Tie” available. Prepending an apology to one copy, and

changing nothing else, led most of eleven judges to abandon the tie and prefer the apologetic copy, in one judge's case nearly always and across judges up to 98 percent; Command R alone disfavoured the apology. A tie is a neutral label that requires affirmative evidence of equality, and a surface addition that one side can claim moved judges off it, in a direction the judge's learned association set. That is fabrication in the anchor's sense, with no corruption at all, and its sign varied by judge.

6.4 What the general robustness literature measures, and what it leaves out

Most of what is known about language models and noisy input is measured as lost accuracy, which has no direction. Synthetic and natural character noise both broke neural translation systems on text people read without difficulty (Belinkov and Bisk, 2018); a single adversarially chosen character lowered a fine-tuned BERT sentiment classifier from 90.3 to 45.8 percent accuracy (Pruthi et al., 2019); character- and word-level perturbations meant to mimic everyday noise degraded several pretrained encoders across tasks (Moradi and Samwald, 2021); adversarial prompts crafted to resemble plausible typos and synonym swaps degraded models across eight tasks (Zhu et al., 2023). For current models, a keyboard-aware typo generator degraded eighteen open models across five tasks in several languages, least on natural language inference, and instruction tuning improved clean accuracy while possibly increasing brittleness (Zhao, R. et al., 2025); a single adversarial character edit lowered Mistral-7B-Instruct's GSM8K accuracy from 43.7 to 38.6 percent and eight edits to 19.2 (Gan et al., 2024). For judges specifically, a stress-testing harness found formatting perturbations more damaging than semantic ones and observed that, on agentic transcripts, some judges missed inserted violations while others over-penalised corrected ones (Dev et al., 2026), and a broad robustness assessment of LLM judges found them manipulable by a range of attacks (Li, S. et al., 2025). An analysis of the whole safety-evaluation pipeline lists judge noise among several sources that can make attacks and defences impossible to compare (Beyer et al., 2025).

What these studies share is a scoring rule that counts a wrong answer the same whichever way it is wrong. For a task with a neutral label, that rule hides exactly the quantity the anchor isolated. The anchor's contribution to this literature is to split the loss by direction; this paper's addition is that the split it found is a fact about where its judges fall, and that the same corruption in a task whose positive label demands evidence is recorded falling the other way. The judge-bias literature has catalogued many directional tendencies, among them position, verbosity, self-preference and leniency (Zheng, L. et al., 2023; Ye et al., 2024; Thakur et al., 2024), but a tendency toward committing rather than withholding has not, to the author's knowledge, been measured as a property of a judge, separately from the content it is

shown. The abstention literature suggests it would be worth measuring: across twenty models, declining to answer unanswerable or underspecified questions remained unsolved, scale helped little, and reasoning fine-tuning lowered abstention by 24 percent on average (Kirichenko et al., 2025), and the survey literature treats abstention as a capability that has to be built and evaluated rather than assumed (Wen et al., 2024). A judge trained toward answering is a judge whose default, when the text gives out, is plausibly a direction.

7. Does noise make bias look larger? It depends on the score

The anchor's conclusion is about a measured quantity: bias measured on noisy text is overestimated. Which quantity a pipeline computes from a judge's labels decides whether that follows.

7.1 Three families of score

A pipeline that labels responses A, B or None can report at least three kinds of number. The first is a commitment rate: the share of responses judged to support either candidate, read as the share that show bias. The second is a signed score: the share judged to support the stereotyped candidate minus the share judged to support the other. The third is a conditional score: among responses judged to commit, the share that commit to the stereotyped candidate. BBQ defines the second and third. Its bias score "reflects the percent of non-UNKNOWN outputs that align with a social bias", computed in disambiguated contexts as $s_{DIS} = 2(n_{biased} / n_{non-UNKNOWN}) - 1$, and in ambiguous contexts it is scaled by error, $s_{AMB} = (1 - accuracy)s_{DIS}$, so that "a biased answer is more harmful if it happens more often" (Parrish et al., 2022). A generation-based adaptation of BBQ measures the probability of neutral and biased continuations (Jin et al., 2025), and any pipeline in which a judge maps free text onto the answer options, as the anchor's does, inherits the structure.

7.2 A derivation

In an ambiguous context the correct answer is UNKNOWN, so accuracy is the share of UNKNOWN answers, U. Write S for the share of all answers that pick the stereotyped candidate and C for the share that pick the other, so that $S + C = 1 - U$ and the biased share among non-UNKNOWN answers is $S / (1 - U)$. Substituting into the definition gives $s_{AMB} = (1 - U)(2S / (1 - U) - 1) = 2S - (1 - U) = S - C$. The ambiguous-context bias score is simply the net stereotype lean over all items.

Now let noise move a share d of all items from UNKNOWN to a candidate, with a fraction q of those moves going to the stereotyped candidate, and a share e from a candidate to UNKNOWN, with a fraction r of those leaving the stereotyped candidate. Then S rises by

qd and falls by re, C rises by $(1 - q)d$ and falls by $(1 - r)e$, and

$$\text{change in } s_{\text{AMB}} = d(2q - 1) - e(2r - 1).$$

Noise can also reverse a verdict from one candidate to the other, and the anchor reports that such reversals make up 39 percent of its fragile judge's flips under combined noise and 16 to 19 percent of the other judges'. Let a share v_{CS} of all items move from the non-stereotyped to the stereotyped candidate and a share v_{SC} move the reverse way. Each such move changes S - C by 2, so the full change is

$$\text{change in } s_{\text{AMB}} = d(2q - 1) - e(2r - 1) + 2(v_{\text{CS}} - v_{\text{SC}}).$$

The commitment rate, $1 - U$, changes by $d - e$; reversals leave it unchanged. The conditional score $s_{\text{DIS}} = 2S / (S + C) - 1$ moves toward $2q - 1$ as d grows.

Three consequences follow, each flat because it is algebra. Fabrication always raises the commitment rate. Fabrication raises the signed score only to the extent that the fabricated commitments favour the stereotype, $q > 1/2$; fabrication that is indifferent to the stereotype, $q = 1/2$, adds nothing to s_{AMB} however large d is; fabrication that runs against the stereotype lowers it. And because s_{DIS} is pulled toward $2q - 1$, indifferent fabrication pulls a nonzero conditional score toward zero, so on that score noise can make a biased judge's measurement look less biased.

7.3 What q is in the anchor

The anchor did not measure q . Its limitations section says so directly: the metric "does not verify whether the newly selected option is stereotype-aligned or counter-stereotypical", the options differ only by name, the option codes "record position rather than any demographic attribute", and for age, disability, socioeconomic status and sexual orientation "nothing in a pair of names identifies which candidate the stereotype would target." For the fragile judge, 258 of 288 fabrications under combined noise went to the second-listed candidate. If the stereotyped candidate, wherever the item defines one, sits second in about half of the items, those fabrications have q near one half and move a signed score little; if it sits second more often, they inflate it; if less, they deflate it. Nothing in the paper fixes that proportion, and for four of seven axes the design does not define a stereotyped candidate at all.

The licensed statement is therefore narrower than the anchor's. Bias measured on noisy text as a rate of directional verdicts was overestimated by the fragile judges, by amounts Table 1 gives. Bias measured as a signed or conditional stereotype score may have been inflated, unchanged or reduced, and the anchor's data cannot say which. The distinction matters for the fairness reading the anchor gives its category results: a category

with many fabrications is a category whose commitment rate is unstable under noise, not necessarily one whose stereotype score is inflated.

7.4 The general lesson for bias measurement

That bias scores are fragile to choices that should not matter is well documented. Semantically equivalent rewordings of templates move bias measurements and the conclusions drawn from them (Seshadri et al., 2022); innocuous changes to how a benchmark is constructed change the degree of bias measured on two benchmarks (Selvam et al., 2022); open-ended generation measurements contradict each other under different prompt sets, metrics and tools (Akyurek et al., 2022); many bias metrics for pretrained models are mutually incompatible and depend heavily on templates and seed words (Delobelle et al., 2021); generation-based and question-answering evaluations of the same models disagree (Jin et al., 2025); and a retrained toxicity API reordered the models on a published leaderboard when the same outputs were rescored (Pozzobon et al., 2023). Classifier-based measures of regard and toxicity have long stood in for bias in generated text (Sheng et al., 2019; Dhamala et al., 2021). The measurement-modelling view treats each such score as an operationalisation whose reliability and validity have to be argued, not assumed (Jacobs and Wallach, 2021), critical review finds the quantitative techniques poorly matched to the harms that motivate them (Blodgett et al., 2020), an inventory of benchmark pitfalls finds the same at the level of individual items (Blodgett et al., 2021), and a recent survey organises the resulting metrics by the level of the model at which they operate (Gallegos et al., 2024). The anchor's contribution sits naturally in this record as a new source of instability, input quality. The derivation above adds that the direction of the resulting error depends on the score's sign structure, which is a choice made after the judge has spoken.

8. Robustness removes the distortion: what the robust judges can bear

8.1 The claim

The anchor's third contribution is that "as a judge becomes more robust the asymmetry diminishes toward parity rather than reversing: robustness mitigates the distortion, it does not invert it." It rests on the two robust judges. Across 100 judge-condition pairs, erasure exceeded fabrication in eight, six of them for GPT-5.4 and two for Gemma, and none significantly.

8.2 Power

Absence of a significant reversal is evidence against reversal only if a reversal of interest could have been detected. Under combined noise at $p = 0.3$, GPT-5.4 produced 28 fabrications and 38 erasures, 66 in all; by the author's arithmetic the two-sided binomial p -value

for that split is about 0.27. With 66 flips, a true erasure-to-fabrication ratio of 1.5 to 1 would be detected at the 5 percent level about a third of the time, and a ratio near 1.9 to 1 about two times in three; under typo noise, with 54 flips, the corresponding powers are lower. Underpowered comparisons of this size are common in NLP evaluation, and they make it difficult to tell noise from a real difference (Card et al., 2020). The data are consistent with parity in GPT-5.4, and also consistent with a moderate reversal; what they rule out is only a strong one.

Gemma's evidence is mixed rather than null. Under typo noise its 31 to 9 split is significant, as the anchor reports; under combined noise its 36 to 22 split is not (p about 0.09 by the author's arithmetic). Pooling the two robust judges gives 61 to 33 under typo, an asymmetry in the anchor's direction, and 64 to 60 under combined, which is parity. The pooled result the anchor reports for all 25 conditions is dominated by the fragile judges: by the author's arithmetic, 85 percent of pooled fabrications under typo at $p = 0.3$ and 88 percent under combined came from Llama and Qwen.

8.3 The floor

For GPT-5.4 a further term enters. Its flip rates at $p = 0.3$ lie between 1.7 and 2.1 percent across the five noise types, a narrow band that the anchor reads as the noise-type hierarchy vanishing once a judge is robust. A judge whose clean verdicts change on rerun at a rate of that order would produce exactly such a band, whatever the noise type, and the studies in Section 3.3 show rerun variation of that order in hosted models run under nominally deterministic settings. Whether GPT-5.4's flips are a response to noise at all is not known, because the clean-versus-clean rerun that would set its floor was not reported. If they are mostly rerun variation, their near-parity says nothing about robustness to noise; it says the judge's own instability is roughly symmetric between None and a direction.

The licensed statement is that in the two fragile judges the asymmetry is large and in the anchor's direction; in Gemma it is present under some conditions and not others; in GPT-5.4 it is not distinguishable from parity, and the anchor's data cannot distinguish parity from a moderate reversal or a response to noise from rerun variation.

9. Measuring through a noisy judge

9.1 A bias rate is a prevalence

A pipeline that reports the share of responses that are biased is estimating a prevalence through an imperfect instrument, and the statistics of that problem are old. If a test has sensitivity Se and specificity Sp , the apparent prevalence p_{obs} relates to the true prevalence p by $p_{obs} = p Se + (1 - p)(1 - Sp)$, and the estimator of Rogan and Gladen (1978) inverts it: $p = (p_{obs} + Sp -$

$1) / (Se + Sp - 1)$. Quantification research arrived at the same correction from the machine-learning side, starting from the observation that unadjusted classify-and-count is often a biased estimator of class prevalence and adjusting it by the classifier's true- and false-positive rates (Forman, 2008); a later reassessment found that classify-and-count variants, once properly optimised, come close to the state of the art though some newer methods still beat them (Moreo and Sebastiani, 2021). The label-noise literature distinguishes noise whose rate depends on the true class from noise that does not, and treats the first as the harder case (Frenay and Verleysen, 2014). If its clean verdicts are taken as correct, a judge whose fabrication rate under corruption exceeds its erasure rate is an instrument whose specificity falls under noise faster than its sensitivity: class-dependent noise in the measurement.

9.2 What the anchor's design gives and withholds

The anchor's self-baseline gives the change in each judge's verdicts but not the change in either error rate, because it compares the judge with itself rather than with a reference label. It does so deliberately, so as not to entangle noise-induced change with differences in clean tendency, and for authentic user text it argues that stance labels would be contested enough to be indistinguishable from the flips being measured. That argument is sound as a reason not to use a single shared reference, and it is also the reason the anchor cannot say how much a measured bias rate on noisy text is off by. A fabrication on an item whose clean None was itself wrong is a correction; the anchor's authors note this skeptical reading and argue against it on the grounds that the rules add no content and the effect scales with corruption and fragility, both of which are fair, and both of which concern the average rather than the item.

Two families of correction need only a small reference sample per noise stratum. Rogan-Gladen-style estimators plug estimated Se and Sp into the inversion above; prediction-powered inference uses a small gold-labelled set to correct the bias of model predictions with valid confidence intervals and makes no assumption about the predictor (Angelopoulos et al., 2023), and the same idea applied to evaluation with model-generated labels increased effective human sample size by up to 50 percent in one study (Boyeau et al., 2024). Chen et al. (2026) compare the two families for LLM-judge evaluation, derive efficient estimators that unify them, and characterise when the prediction-powered form has strictly smaller variance. For a bias pipeline on noisy text, the implication is that a modest double-annotated sample of noisy items, stratified by noise level, turns a flip analysis into a corrected estimate with an interval; contested annotation widens the interval but does not prevent it.

9.3 Why the direction must be measured per judge

The prevalence view also says why the direction cannot be imported. The bias in p_{obs} is $(1 - p)(1 - Sp) - p(1 - Se)$, and its sign depends on the base rate and on both error rates, each of which a judge's prompt, default and response to corruption can move independently. The anchor's fragile judges lose specificity under noise; the three safety judges in Figure 1 lose sensitivity under restyling; LLaMA Guard loses specificity. A pipeline that learned the direction from one of these studies and assumed it for its own judge could correct in the wrong direction.

10. A protocol

Algorithm 1 states the measurements that would let a bias pipeline say which way noise moves its judge and how much its reported number is off, given the published methods each step borrows. It proposes no new score.

Steps S1 to S4 need no labels and no new data beyond the judge's own outputs; S2 and S3 are the calibration and position-balancing controls the label-bias and judge-bias literatures already use (Zhao, T. et al., 2021; Fei et al., 2023; Wang, P. et al., 2023; Zheng, C. et al., 2023). S5 and S6 each need one extra run at one noise level. S7 needs the demographic coding the anchor's limitations section describes as possible for three of its axes. S8 is the only step that needs people, and Section 9.2 gives the reason its cost is modest. Step S9 is the one the anchor's abstract does not observe.

11. Measurements that would settle the open part

Table 3 sets out what each account in Section 5 predicts under the controls of Algorithm 1. It is the author's derivation from the accounts as stated; a dash means the account makes no definite prediction.

Three of these are cheap enough to run on the anchor's released items and judge script. The released benchmark carries the two candidate names, their position codes, the reference stance label and the response for each of the 3,822 items, and the released noise toolkit regenerates every condition from a seed. Running the fragile judge on content-free inputs and on both option orders would say how much of its 258 to 30 split is position; running it at one noise level with negation tokens protected would say how much of the asymmetry is the loss of None's evidence. A fourth, re-annotating a stratified sample of noisy items, is the one that would turn the anchor's flip counts into an estimate of how far a noisy bias rate is off.

12. Objections

The base rate is balanced, so the asymmetry is real. It is real, and Section 3.1 agrees. The argument of this paper is not that the asymmetry is an artefact but that its

direction belongs to a judge and a prompt, and that its consequence for a bias score depends on the score.

The direction held in every condition where it was significant. It held, where significant, for three judges, under one prompt and one response set. A uniform direction across judges sharing one prompt is what the definitions account predicts; it does not distinguish that account from integration. The harm-detection record, where the evidence-demanding label is the positive one, shows the opposite direction under comparable corruption, and the restyling study shows both directions under one change.

Harm detection is a different task. It is a different task with the same structure: a classifier with one label that demands evidence and one that does not, applied to text whose surface was changed without changing what a person would say about it. Section 6 uses it for that structure and for nothing about the anchor's judges. Where the comparison is weak, as with adversarial rows whose direction the attacker chose, Section 6.1 says so.

Corruption cannot surface bias that was not there. The anchor argues that its rules add no content, so a newly committed verdict cannot reflect information the noise revealed. That is right for content and incomplete for evidence. Corruption that damages a negation or a statement of insufficiency does not add a lean; it removes the sentence that cancelled one. A verdict that changes because a hedge became unreadable is a measurement failure, as the anchor says, but its direction is set by where the hedges were, which is a fact about the responses as much as about the judge.

A judge that fabricates is the dangerous case, whatever the cause. For a pipeline that reports a rate of directional verdicts, yes, and the anchor's warning is well aimed at it. For a pipeline that reports a signed stereotype score, the danger is not established, and a correction that assumed inflation could remove bias that was there. Section 7 gives the condition.

13. What is not known

The content-free output distribution of any of the anchor's four judges is not known. The clean-versus-clean rerun rate of GPT-5.4 on these items is not known. The share of the anchor's fabrications that land on a stereotyped candidate is not known for any axis. How the asymmetry responds to a prompt in which None is the default is not known. How it responds to corruption that spares negation, or spares names, is not known. Whether informal register is itself read as a lean by bias judges, as dialect is read as offence by toxicity classifiers, is not known. How far a bias rate measured on noisy text differs from the rate a reference annotation would give is not known for any judge. Whether any of the anchor's findings hold on authentic noisy user text, rather than rule-based noise on model-generated responses, is not known; the anchor names this as its next question.

Algorithm 1. Direction-aware bias measurement on noisy text.

Input:	judge J with prompt P; items x with reference labels y on a small sample; noise family N at levels s; bias score F
Output:	per-level corrected estimate of F, and the direction of the noise effect for this J and P
S1:	floor. Run J twice on clean x. Record the clean-vs-clean transition matrix T0. Every later flip count is read against T0.
S2:	default. Run J on content-free inputs and on corpus-random words in place of x. Record J's output distribution D0. If D0 puts most mass on a candidate: expect fabrication. If D0 puts most mass on None: expect erasure.
S3:	order. Present every item in both option orders. Report flips by position; average verdicts over orders before scoring.
S4:	sweep. For each level s, record the full transition matrix T(s). Report fabrication and erasure as rates of their pools, not only as counts, minus the corresponding rates in T0.
S5:	cue. Repeat one level with negation and hedge tokens protected, and once with names protected. A drop in fabrication when cues are protected assigns it to evidence removal.
S6:	definitions. Repeat one level with P reworded so that None is the default. A sign change assigns the direction to P.
S7:	alignment. For signed or conditional F, record for each fabrication, erasure and reversal whether it lands on or leaves the stereotyped candidate. Report q, r, v_CS and v_SC; the change in a signed score is $d(2q - 1) - e(2r - 1) + 2(v_{CS} - v_{SC})$.
S8:	correct. Annotate a stratified sample at each level. Estimate J's error rates per level and report F with an interval by a Rogan-Gladen or prediction-powered estimator.
S9:	scope. State the direction for J, P, N and F only. Do not transfer it to another judge, prompt, noise family or score.

Table 3. Predictions of each account of the anchor's asymmetry under each control.

Control	Integration	Definitions	Default	Cue removal
None-default prompt (S6)	asymmetry stays	shrinks or reverses	-	stays
Content-free output (S2)	-	-	predicts destination	-
Both option orders (S3)	stays	stays	position share vanishes	stays
Negation protected (S5)	shrinks a little	stays	stays	shrinks sharply
Names protected (S5)	shrinks a little	stays	stays	erasure falls
Judge with None default	fabricates	fabricates under P	erases	fabricates

14. Conclusion

A judge reading a corrupted passage has less to go on and must still answer. Where its answer goes depends on what its instructions make hard to earn, on what it says when the passage says little, and on which words the corruption happened to strike. In the anchor all three point the same way: the neutral label demands affirmative evidence, the evidence for it plausibly lives in negations that typos damage, and the most fragile judge falls toward the second option it already favoured. The result is a large, real, base-rate-robust shift from neutral to directional verdicts. In harm detection the three point the other way, and corruption erases. One restyling study shows a single change sending different judges in different directions. The anchor's measurement establishes that its fragile judges overstate the rate of directional verdicts on noisy text. It does not establish that bias scores built on those verdicts are inflated, because the signed score moves by the net stereotype alignment of its fabrications, erasures and reversals, none of which the anchor measured; nor that robustness removes the distortion without reversing it, because the closed judge's evidence is underpowered and has no

floor. Four cheap controls, a content-free run, both option orders, a cue-protected noise condition and a reworded prompt, would assign the direction for any judge, and a small stratified annotation would say how far its numbers are off. Until those are reported, the direction of noise in a bias judge is a property to be measured, not a fact to be assumed.

References

- Akyurek, A. F., Kocyigit, M. Y., Paik, S., & Wijaya, D. (2022). Challenges in Measuring Bias via Open-Ended Language Generation. arXiv:2205.11601. doi:10.48550/arXiv.2205.11601.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-Powered Inference. arXiv:2301.09633. doi:10.48550/arXiv.2301.09633.
- Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., et al. (2024). Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667. doi:10.48550/arXiv.2408.04667.
- Belinkov, Y., & Bisk, Y. (2018). Synthetic and Natural Noise Both Break Neural Machine Translation. ICLR 2018. arXiv:1711.02173. doi:10.48550/arXiv.1711.02173.
- Beyer, T., Xhonneux, S., Geisler, S., Gidel, G., Schwinn, L., & Gunnemann, S. (2025). LLM-Safety Evaluations Lack Robustness. arXiv:2503.02574. doi:10.48550/arXiv.2503.02574.
- Blodgett, S. L., Barocas, S., Daume III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in

- NLP. arXiv:2005.14050. doi:10.48550/arXiv.2005.14050.
- Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *Proceedings of ACL-IJCNLP 2021*. doi:10.18653/v1/2021.acl-long.81.
- Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. arXiv:1903.04561. doi:10.48550/arXiv.1903.04561.
- Boyeau, P., Angelopoulos, A. N., Yosef, N., Malik, J., & Jordan, M. I. (2024). AutoEval Done Right: Using Synthetic Data for Model Evaluation. arXiv:2403.07008. doi:10.48550/arXiv.2403.07008.
- Cantini, R., Orsino, A., Ruggiero, M., & Talia, D. (2025). Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. *Machine Learning*, 114(11), 249. doi:10.1007/s10994-025-06862-6.
- Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., & Jurafsky, D. (2020). With Little Power Comes Great Responsibility. arXiv:2010.06595. doi:10.48550/arXiv.2010.06595.
- Chen, H., & Goldfarb-Tarrant, S. (2025). Safer or Luckier? LLMs as Safety Evaluators Are Not Robust to Artifacts. arXiv:2503.09347. doi:10.48550/arXiv.2503.09347.
- Chen, Y. T., Lu, S., Li, S., Guo, M., & Li, S. (2026). Efficient Inference for Noisy LLM-as-a-Judge Evaluation. arXiv:2601.05420. doi:10.48550/arXiv.2601.05420.
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. arXiv:1905.12516. doi:10.48550/arXiv.1905.12516.
- Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2021). Measuring Fairness with Biased Rulers: A Survey on Quantifying Biases in Pretrained Language Models. arXiv:2112.07447. doi:10.48550/arXiv.2112.07447.
- Dev, S., Sloan, A., Kavner, J., Kong, N., & Sandler, M. (2026). Judge Reliability Harness: Stress Testing the Reliability of LLM Judges. arXiv:2603.05399. doi:10.48550/arXiv.2603.05399.
- Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. arXiv:2101.11718. doi:10.48550/arXiv.2101.11718.
- Eiras, F., Zemour, E., Lin, E., & Mugunthan, V. (2025). Know Thy Judge: On the Robustness Meta-Evaluation of LLM Safety Judges. arXiv:2503.04474. doi:10.48550/arXiv.2503.04474.
- Fei, Y., Hou, Y., Chen, Z., & Bosselut, A. (2023). Mitigating Label Biases for In-context Learning. arXiv:2305.19148. doi:10.48550/arXiv.2305.19148.
- Forman, G. (2008). Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-008-0097-y.
- Frenay, B., & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*. doi:10.1109/TNNLS.2013.2292894.
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., et al. (2024). Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770. doi:10.48550/arXiv.2309.00770.
- Gan, E., Zhao, Y., Cheng, L., Mao, Y., Goyal, A., Kawaguchi, K., Kan, M.-Y., & Shieh, M. (2024). Reasoning Robustness of LLMs to Adversarial Typographical Errors. arXiv:2411.05345. doi:10.48550/arXiv.2411.05345.
- Gond, R., Kamath, A. K., Ramjee, R., & Panwar, A. (2026). LLM-42: Enabling Determinism in LLM Inference with Verified Speculation. arXiv:2601.17768. doi:10.48550/arXiv.2601.17768.
- Grondahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All You Need is "Love": Evading Hate-speech Detection. arXiv:1808.09115. doi:10.48550/arXiv.1808.09115.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., et al. (2024). A Survey on LLM-as-a-Judge. arXiv:2411.15594. doi:10.48550/arXiv.2411.15594.
- Hackett, W., Birch, L., Trawicki, S., Suri, N., & Garraghan, P. (2025). Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems. arXiv:2504.11168. doi:10.48550/arXiv.2504.11168.
- 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. arXiv:2203.09509. doi:10.48550/arXiv.2203.09509.
- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, 633, 147-154. doi:10.1038/s41586-024-07856-5.
- Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordani, A., & Courville, A. (2021). Understanding by Understanding Not: Modeling Negation in Language Models. arXiv:2105.03519. doi:10.48550/arXiv.2105.03519.
- Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138. doi:10.48550/arXiv.1702.08138.
- Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of FAccT 2021*. arXiv:1912.05511. doi:10.48550/arXiv.1912.05511.
- Jin, J., Kang, W., Myung, J., & Oh, A. (2025). Social Bias Benchmark for Generation: A Comparison of Generation and QA-Based Evaluations. *Findings of ACL 2025*. arXiv:2503.06987. doi:10.48550/arXiv.2503.06987.
- Kassner, N., & Schutze, H. (2020). Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. arXiv:1911.03343. doi:10.48550/arXiv.1911.03343.
- Kirichenko, P., Ibrahim, M., Chaudhuri, K., & Bell, S. J. (2025). AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. arXiv:2506.09038. doi:10.48550/arXiv.2506.09038.
- Kurita, K., Belova, A., & Anastasopoulos, A. (2019). Towards Robust Toxic Content Classification. arXiv:1912.06872. doi:10.48550/arXiv.1912.06872.
- Li, S., Xu, C., Wang, J., Gong, X., Chen, C., Zhang, J., Wang, J., Lam, K.-Y., & Ji, S. (2025). LLMs Cannot Reliably Judge (Yet?): A Comprehensive Assessment on the Robustness of LLM-as-a-Judge. arXiv:2506.09443. doi:10.48550/arXiv.2506.09443.
- Lim, Y., Choi, H., & Kim, M. (2026). Reliable to Expressive: A Curriculum for Rubric-Following Safety Judges. arXiv:2606.09165. doi:10.48550/arXiv.2606.09165.
- Moradi, M., & Samwald, M. (2021). Evaluating the Robustness of Neural Language Models to Input Perturbations. arXiv:2108.12237. doi:10.48550/arXiv.2108.12237.
- Moreo, A., & Sebastiani, F. (2021). Re-Assessing the Classify and Count Quantification Method. arXiv:2011.02552. doi:10.48550/arXiv.2011.02552.
- Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023). An Empirical Study of the Non-determinism of ChatGPT in Code Generation. arXiv:2308.02828. doi:10.48550/arXiv.2308.02828.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering. *Findings of ACL 2022*. arXiv:2110.08193. doi:10.48550/arXiv.2110.08193.
- Pezeshkpour, P., & Hruschka, E. (2023). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv:2308.11483. doi:10.48550/arXiv.2308.11483.
- Pozzobon, L., Ermis, B., Lewis, P., & Hooker, S. (2023). On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research. arXiv:2304.12397. doi:10.48550/arXiv.2304.12397.
- Pruthi, D., Dhingra, B., & Lipton, Z. C. (2019). Combating Adversarial Misspellings with Robust Word Recognition. arXiv:1905.11268. doi:10.48550/arXiv.1905.11268.
- Raina, V., Liusie, A., & Gales, M. (2024). Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM

- Assessment. arXiv:2402.14016. doi:10.48550/arXiv.2402.14016.
- Resende, G. H., Nery, L. F., Benevenuto, F., Zannettou, S., & Figueiredo, F. (2024). A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions. arXiv:2401.12720. doi:10.48550/arXiv.2401.12720.
- Rogan, W. J., & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*. doi:10.1093/oxfordjournals.aje.a112510.
- Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv:2308.01263. doi:10.48550/arXiv.2308.01263.
- Rottger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of ACL-IJCNLP 2021*. arXiv:2012.15606. doi:10.48550/arXiv.2012.15606.
- Ryu, D., Lee, J., Hwang, Y., & Bak, J. (2026). When Noise Fabricates Bias: The Fragility of LLM-as-a-Judge Bias Measurement under Noisy Text. *W-NUT 2026* (accepted). arXiv:2609.11067. doi:10.48550/arXiv.2609.11067.
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of ACL 2019*. doi:10.18653/v1/P19-1163.
- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324.
- Selvam, N. R., Dev, S., Khashabi, D., Khot, T., & Chang, K.-W. (2022). The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks. arXiv:2210.10040. doi:10.48550/arXiv.2210.10040.
- Seshadri, P., Pezeshkpour, P., & Singh, S. (2022). Quantifying Social Biases Using Templates is Unreliable. arXiv:2210.04337. doi:10.48550/arXiv.2210.04337.
- Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. arXiv:1909.01326. doi:10.48550/arXiv.1909.01326.
- Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., & Vosoughi, S. (2024). Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. arXiv:2406.07791. doi:10.48550/arXiv.2406.07791.
- Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724. doi:10.48550/arXiv.2405.01724.
- Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2024). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. arXiv:2406.12624. doi:10.48550/arXiv.2406.12624.
- Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. arXiv:2306.08189. doi:10.48550/arXiv.2306.08189.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926. doi:10.48550/arXiv.2305.17926.
- Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., & Wang, L. L. (2024). Know Your Limits: A Survey of Abstinence in Large Language Models. arXiv:2407.18418. doi:10.48550/arXiv.2407.18418.
- Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting Racial Bias in Hate Speech Detection. arXiv:2005.12246. doi:10.48550/arXiv.2005.12246.
- Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., et al. (2024). Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736. doi:10.48550/arXiv.2410.02736.
- Zhao, R., Liu, Y., Altinger, L., Schutze, H., & Hedderich, M. A. (2025). Evaluating Robustness of Large Language Models Against Multilingual Typographical Errors. arXiv:2510.09536. doi:10.48550/arXiv.2510.09536.
- Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690. doi:10.48550/arXiv.2102.09690.
- Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2023). Large Language Models Are Not Robust Multiple Choice Selectors. arXiv:2309.03882. doi:10.48550/arXiv.2309.03882.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.
- Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., et al. (2023). PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv:2306.04528. doi:10.48550/arXiv.2306.04528.
- Zhou, H., Wan, X., Proleev, L., Mincu, D., Heller, K., & Roy, S. (2023). Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. arXiv:2309.17249. doi:10.48550/arXiv.2309.17249.