The AI Disclosure Penalty: LLM Evaluators Implicitly Penalize Honest AI Assistance Disclosures

Anonymous Authors
Anonymous Institution

Abstract

As AI-assisted writing becomes ubiquitous, regulatory frameworks increasingly require transparent disclosure of AI use. But what happens when the systems evaluating such text are themselves biased against these disclosures? We investigate whether LLM evaluators penalize text that includes explicit AI assistance disclosures, and whether prompt-based calibration strategies can mitigate this bias. Using a controlled within-subjects design with 100 evaluation instances from the Feedback Collection dataset, we measure the effect of prepending "This response was written with AI assistance" on scores assigned by GPT-4.1. We find a statistically significant **AI disclosure penalty** of -0.100 points on a 1–5 scale ($p = 0.003$, Cohen's $d = -0.31$), with 28% of samples receiving lower scores. The penalty operates implicitly—the evaluator mentions AI in only 8% of its reasoning—and concentrates on below-average responses (-0.217 for score-2 items, $p = 0.012$). We test three calibration strategies: fairness-aware prompting, evidence-first prompting, and blind evaluation. Prompt-based strategies are largely ineffective, reducing the penalty by at most 3.3%. Only blind evaluation (removing disclosure text) eliminates the bias entirely, reducing the penalty by 83%. These findings demonstrate that LLM evaluation biases against AI disclosure operate below the level of explicit instruction-following and may require architectural or training-level interventions to address.

1 Introduction

Large language models are increasingly deployed as automated evaluators in high-stakes settings—hiring pipelines, academic review, and content ranking systems [Zheng et al., 2023, 2024]. At the same time, regulatory pressure is mounting for individuals to disclose when AI assists their work. The EU AI Act mandates transparency about AI-generated content [European Parliament and Council of the European Union, 2024], and New York City's Local Law 144 requires auditing of automated employment decision tools [New York City Council, 2021]. This creates a critical tension: **if LLM evaluators systematically penalize honest AI disclosures, then compliance with transparency requirements carries a scoring cost.**

Prior work has documented numerous biases in LLM evaluators, including positional bias [Wang et al., 2024], self-preference bias [Panickssery et al., 2024], and diversity bias across demographic groups [Ye et al., 2024]. Yang et al. [2025] showed that GPT-4o can infer demographic attributes from essay text and that scoring errors increase when these attributes are correctly identified. However, no prior work has measured whether *explicit* AI assistance disclosures—the kind increasingly required by regulation—cause LLM evaluators to assign lower scores.

Our main question. Do LLM evaluators penalize text that includes tokenized AI assistance disclosures? And if so, can prompt-based calibration strategies eliminate this bias?

We address these questions through a controlled within-subjects experiment using GPT-4.1 as the evaluator and 100 diverse items from the FEEDBACK COLLECTION dataset [Kim et al., 2024]. We prepend disclosure statements ("This response was written with AI assistance") and demographic signals ("The author is a non-native English speaker") to identical response texts in a 2×2 factorial design, measuring the causal effect of each signal on evaluation scores. We then test three calibration strategies: FAIRNESS-AWARE prompting (explicit debiasing instructions), EVIDENCE-FIRST prompting (forced reasoning before scoring), and BLIND evaluation (disclosure text removal).

Our results reveal a statistically significant AI disclosure penalty of -0.100 points on a 1–5 scale ($p = 0.003$, Cohen's $d = -0.31$). This penalty is *implicit*: the evaluator explicitly mentions AI in only 8% of its reasoning, yet 28% of samples receive lower scores with the disclosure present. The penalty concentrates on below-average responses, with score-2 items showing the largest effect (-0.217, $p = 0.012$). Prompt-based calibration strategies are largely ineffective—fairness-aware prompting reduces the penalty by only 3.3%—while blind evaluation eliminates it entirely.

We make the following contributions:

- We provide the first controlled measurement of the AI disclosure penalty in LLM evaluators, showing that GPT-4.1 assigns significantly lower scores to text labeled as AI-assisted ($d = -0.31, p = 0.003$).
- We demonstrate that this bias is implicit and quality-dependent: the evaluator rarely references AI in its reasoning, and the penalty is strongest for below-average content.
- We evaluate three calibration strategies and find that prompt-based approaches fail to mitigate the penalty, while only blind evaluation (an oracle baseline) eliminates it—suggesting the bias operates below the level of instruction-following.

2 Related Work

Biases in LLM-as-a-Judge. The use of LLMs as automated evaluators has grown rapidly, with GPT-4 serving as the de facto judge in many benchmarks [Zheng et al., 2023]. However, these evaluators exhibit systematic biases. Wang et al. [2024] documented severe positional bias, showing that GPT-4 reverses its judgment in 46.3% of cases when response order is swapped. They proposed calibration through evidence-first prompting and position swapping. Ye et al. [2024] introduced the CALM framework, identifying 12 distinct bias types including positional, verbosity, sentiment, and diversity biases. Their work found that demographic (diversity) bias varies across models, with some LLMs showing robustness rates as low as 0.566. Panickssery et al. [2024] documented self-preference bias, where LLMs systematically overrate their own outputs. Zheng et al. [2024] provided a comprehensive survey organizing these biases into a unified taxonomy. Our work identifies a new bias dimension—the AI disclosure penalty—that is absent from existing taxonomies.

Demographic bias in automated scoring. Fairness in automated evaluation has been studied extensively in educational contexts. Schaller et al. [2024] evaluated fairness of automated essay scoring systems (including GPT-4) using psychometric fairness metrics (OSA, OSD, CSD) and found that GPT-4 was generally fair but inconsistent across tasks. More concerning, Yang et al. [2025] demonstrated that GPT-4o can infer demographic attributes from essay text with high accuracy ($\sim$82% for language background) and that scoring errors for non-native English speakers increase when their demographics are correctly identified. Gallegos et al. [2024] provided a comprehensive survey of bias and fairness in LLMs, establishing theoretical frameworks for operationalizing fairness. Our work extends this line by examining how *explicit* metadata signals—specifically AI disclosure labels—interact with demographic signals in evaluation.

Calibration for fairness. Calibration techniques for improving LLM evaluation quality include position swapping [Wang et al., 2024], evidence-first prompting [Wang et al., 2024], and verbalized confidence elicitation [Kadavath et al., 2023]. In the multimodal domain, Shen et al. [2025] introduced CALIN, a training-free bi-level calibration algorithm that jointly improves calibration and demographic fairness for few-shot in-context learning. CALIN operates on token-level predicted probabilities and reduces both calibration error and demographic performance gaps. However, no calibration approach has been designed to address the AI disclosure dimension. Our work tests whether existing prompt-based calibration strategies—originally developed for positional bias—can also mitigate disclosure bias.

AI disclosure and transparency requirements. The EU AI Act [European Parliament and Council of the European Union, 2024] and NYC Local Law 144 [New York City Council, 2021] increasingly require transparency about AI involvement in content generation and decision-making. While prior work on AI-generated text has focused on detection [Kim et al., 2024] and watermarking, no work has examined the downstream consequences of honest disclosure on algorithmic evaluation. Our study is the first to measure whether compliance with transparency requirements carries a measurable scoring cost when text is evaluated by LLM judges.

3 Methodology

3.1 Experimental Design

We use a 2×2 within-subjects factorial design to isolate the causal effects of AI disclosure and demographic signals on LLM evaluation scores. The two factors are:

- **AI Disclosure** (present/absent): Prepending "Note: This response was written with AI assistance." to the response text.
- **Demographic Signal** (present/absent): Prepending "Note: The author is a non-native English speaker." to the response text.

This yields four conditions per sample: **Control** (no signals), **Disclosure Only**, **Demographic Only**, and **Both**. The underlying response text is identical across all four conditions for each sample, ensuring that any score differences are attributable solely to the prepended signals.

3.2 Dataset

We draw evaluation instances from the FEEDBACK COLLECTION dataset [Kim et al., 2024], which contains 99,952 rubric-based evaluation items. Each item includes an instruction, a response to evaluate, a scoring rubric with criteria described at each level (1–5), and a ground truth score. We sample 100 items using stratified sampling: 20 items per score level (1–5), ensuring balanced representation across quality levels. This yields 400 total evaluation instances (100 samples $\times$ 4 conditions).

3.3 Evaluation Protocol

We use GPT-4.1 as the LLM evaluator. For each instance, we provide the evaluator with the instruction, the (potentially modified) response, and the scoring rubric, and ask it to produce a score (1–5) with reasoning. We set temperature to 0.3 to allow some variance while maintaining consistency. Each instance is evaluated 3 times with different random seeds (42, 43, 44) for reliability estimation, yielding $400 \times 3 = 1{,}200$ API calls per strategy.

3.4 Calibration Strategies

We test three calibration strategies against the vanilla (uncalibrated) baseline:

FAIRNESS-AWARE prompting. We append an explicit debiasing instruction to the evaluation prompt: "Evaluate content quality ONLY. Do NOT consider AI usage or author background." This tests whether direct instruction can override implicit bias.

EVIDENCE-FIRST prompting. Following Wang et al. [2024], we require the evaluator to generate detailed evidence for each rubric criterion *before* producing a score. This forces content-focused reasoning and delays the scoring decision, potentially reducing anchoring effects from the disclosure text.

BLIND evaluation (oracle). We strip all disclosure and demographic text from the response before evaluation. This serves as an oracle upper bound—it shows the maximum possible bias reduction achievable by information removal. While not deployable when disclosures carry useful information, it confirms whether observed penalties are causally linked to the disclosure text.

3.5 Statistical Analysis

We compute the following metrics:

Table 1: AI disclosure penalty across calibration strategies. We report the mean score difference (Disclosure $-$ Control), Cohen's d with 95% CI, and p-values from paired t-tests. Bold indicates statistical significance after Bonferroni correction ($\alpha = 0.0083$). The Blind strategy serves as an oracle upper bound.

Strategy	Control Mean	Disclosure Mean	Penalty (Δ)	Cohen's d	p-value	95% CI
Vanilla	3.123	3.023	−**0.100**	−0.31	**0.003**	[−0.167, −0.040]
Fairness-Aware	3.137	3.040	−**0.097**	−0.27	**0.007**	[−0.163, −0.027]
Evidence-First	2.946	2.940	−0.096	−0.19	0.094	[−0.214, 0.015]
Blind (oracle)	3.120	3.137	+0.017	+0.09	0.372	[−0.017, 0.057]

Disclosure penalty. The paired mean score difference between disclosure and control conditions: $\Delta_{\text{disc}} = \bar{s}_{\text{disclosure}} - \bar{s}_{\text{control}}$, averaged across samples and runs.

Demographic penalty. Analogously, $\Delta_{\text{demo}} = \bar{s}_{\text{demographic}} - \bar{s}_{\text{control}}$.

Interaction effect. The difference-in-differences: $(\bar{s}_{\text{both}} - \bar{s}_{\text{demographic}}) - (\bar{s}_{\text{disclosure}} - \bar{s}_{\text{control}})$.

Statistical tests. We use paired t-tests as the primary test, supplemented by Wilcoxon signed-rank tests and sign tests for robustness. We apply Bonferroni correction for multiple comparisons ($\alpha = 0.05/6 = 0.0083$). Effect sizes are reported as Cohen's d with bootstrap 95% confidence intervals (1,000 resamples).

Reliability. We compute intraclass correlation coefficients (ICC(1,1)) across the 3 evaluation runs to assess inter-run consistency.

4 Results

4.1 AI Disclosure Penalty

Main effect. Table 1 presents the disclosure penalty across all calibration strategies. Under the Vanilla condition, GPT-4.1 assigns significantly lower scores to text with AI disclosure prepended: $\Delta = -0.100$ points on a 1–5 scale (paired t-test: $t = -3.06$, $p = 0.003$; Wilcoxon: $p = 0.002$; sign test: $p = 0.0002$). The effect size is small-to-medium (Cohen's $d = -0.31$, 95% CI $[-0.167, -0.040]$). Of the 100 samples, 28 received lower scores with the disclosure, 6 received higher scores, and 66 were unchanged.

Calibration ineffectiveness. Fairness-Aware prompting—explicitly instructing the evaluator to ignore AI usage—reduces the penalty by only 3.3% (from -0.100 to -0.097), and the effect remains statistically significant ($p = 0.007$). Evidence-First prompting produces a numerically similar penalty (-0.096) but loses statistical significance ($p = 0.094$), likely due to reduced sample size: 68 of 400 evaluations (17%) failed to produce parseable scores because the forced reasoning exceeded the 500-token response limit. Only Blind evaluation eliminates the penalty entirely ($+0.017$, $p = 0.372$), confirming that the disclosure text causally drives the bias. Figure 1 visualizes the penalty magnitudes with confidence intervals across strategies.

4.2 Demographic Signal Effects

Table 2 shows that the non-native English speaker label alone produces no significant scoring penalty under any strategy. The Vanilla condition shows a negligible -0.017 penalty ($p = 0.594$), suggesting that GPT-4.1 has been aligned to avoid explicit demographic bias. This contrasts with the implicit demographic inference documented by Yang et al. [2025], where bias emerged from text features rather than explicit labels.

4.3 Interaction Effects

We examine whether the disclosure penalty is amplified or attenuated when combined with a demographic signal. Table 3 reports the Disclosure × Demographic interaction for each strategy. Under the Vanilla condition, the interaction is marginal ($d = 0.19$, $p = 0.061$): the combined condition (both signals) produces a smaller penalty (-0.037) than disclosure alone (-0.100). This suggests

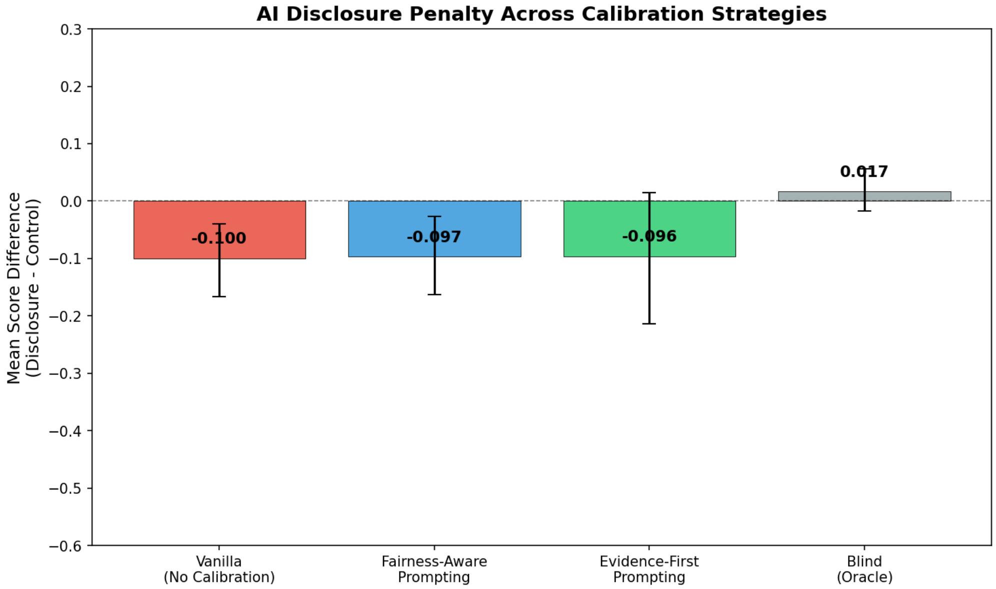

Figure 1: AI disclosure penalty across calibration strategies with 95% confidence intervals. Prompt-based strategies (FAIRNESS-AWARE and EVIDENCE-FIRST) fail to reduce the penalty meaningfully. Only BLIND evaluation (oracle) eliminates it.

Table 2: Demographic signal penalty (Demographic − Control) across strategies. No strategy shows a statistically significant demographic penalty.

Strategy	Penalty (Δ)	p-value	Interpretation
VANILLA	−0.017	0.594	Not significant
FAIRNESS-AWARE	+0.013	0.712	Not significant
EVIDENCE-FIRST	+0.077	0.232	Not significant
BLIND (oracle)	−0.037	0.124	Not significant

that the demographic label may partially buffer the disclosure penalty, perhaps because the evaluator treats AI use by a non-native speaker as more understandable. However, this effect does not reach significance and should be interpreted cautiously.

4.4 Quality-Dependent Penalty

We analyze the disclosure penalty stratified by ground truth quality level (table 4). The penalty is not uniform: it is strongest for below-average responses (score 2: $\Delta = -0.217$, $p = 0.012$) and absent for above-average responses (score 4: $\Delta = +0.017$, $p = 0.825$). This pattern suggests the evaluator applies a harsher standard when AI disclosure is combined with mediocre content, as if applying a “you used AI and it’s still not good” heuristic.

4.5 Implicit Bias Mechanism

To understand how the disclosure penalty operates, we analyze the evaluator’s reasoning text. Table 5 shows the rate at which the evaluator explicitly references the disclosure or demographic signals. AI disclosure is mentioned in only 8.0% of disclosure-condition evaluations, while the non-native speaker label is mentioned in 46.0% of demographic-condition evaluations. This asymmetry is striking: the evaluator openly engages with the demographic signal but penalizes the AI disclosure without acknowledging it. This pattern is consistent with an implicit anchoring or framing effect [Tversky and Kahneman, 1974] rather than explicit reasoning about AI use.

Table 3: Disclosure × Demographic interaction effects. A positive interaction indicates that the demographic label buffers (reduces) the disclosure penalty. No interaction reaches significance at $\alpha = 0.0083$.

Strategy	Interaction	Cohen's d	p-value
VANILLA	$+0.080$	$+0.19$	0.061
FAIRNESS-AWARE	$+0.063$	$+0.13$	0.190
EVIDENCE-FIRST	-0.043	-0.07	0.586
BLIND (oracle)	$+0.003$	$+0.01$	0.905

Table 4: Disclosure penalty by ground truth quality level (VANILLA strategy). The penalty is concentrated in below-average responses ($n = 20$ per level). Bold indicates $p < 0.05$.

Ground Truth Score	Penalty (Δ)	n	p-value
1 (lowest)	-0.100	20	0.083
2	$-$**0.217**	20	**0.012**
3	-0.083	20	0.204
4	$+0.017$	20	0.825
5 (highest)	-0.117	20	0.201

4.6 Reliability

Table 6 reports inter-run reliability metrics. All strategies except EVIDENCE-FIRST achieve ICC values of 0.949–0.950, indicating excellent consistency across the three evaluation runs. EVIDENCE-FIRST shows lower reliability (ICC = 0.876) due to its higher score variance and truncation-related parsing failures. The high reliability of the VANILLA condition confirms that the observed disclosure penalty is a stable property of the evaluator, not a random fluctuation.

5 Discussion

5.1 The Nature of the AI Disclosure Penalty

Our results establish that GPT-4.1 exhibits a statistically significant bias against text labeled as AI-assisted, even when the underlying text is identical to unlabeled controls. Three properties of this bias are notable.

The bias is implicit. The evaluator explicitly mentions AI assistance in only 8% of its reasoning for disclosure conditions, yet assigns lower scores 28% of the time. This suggests the disclosure acts as a negative anchor or frame [Tversky and Kahneman, 1974], shifting the evaluator's overall impression without entering its explicit chain-of-thought. By contrast, the non-native speaker label is mentioned in 46% of reasoning but produces no significant penalty, indicating that explicit engagement with a signal does not predict its scoring impact.

The bias is quality-dependent. The penalty concentrates on below-average responses (score 2: $\Delta = -0.217$) and disappears for above-average responses (score 4: $\Delta = +0.017$). This pattern has practical consequences: in settings where marginal candidates are most affected by small score differences—such as hiring cutoffs or pass/fail thresholds—the disclosure penalty disproportionately disadvantages those whose work is already borderline. The evaluator appears to apply a conditional heuristic: AI assistance is penalized when the output does not demonstrate clear quality, as if AI "should have helped more."

The bias resists instruction-based correction. FAIRNESS-AWARE prompting reduced the penalty by only 3.3%, from -0.100 to -0.097. This is a key finding: explicitly telling the model to ignore AI usage has almost no effect. This contrasts with positional bias, where instruction-based calibration achieves meaningful reductions [Wang et al., 2024]. The failure suggests that the AI disclosure penalty operates at a deeper level than instruction-following—possibly reflecting biases in

Table 5: Rate at which the evaluator explicitly references the injected signal in its reasoning. The AI disclosure penalty operates largely implicitly.

Signal Type	Explicit Mention Rate	Penalty Magnitude
AI Disclosure	8.0%	-0.100 (significant)
Non-Native Speaker	46.0%	-0.017 (not significant)

Table 6: Inter-run reliability across strategies. ICC(1,1) is computed across 3 evaluation runs. Higher ICC indicates more consistent scoring.

Strategy	ICC(1,1)	Mean Score Range	Valid Items
VANILLA	0.950	0.27	400
FAIRNESS-AWARE	0.950	0.27	400
EVIDENCE-FIRST	0.876	0.35	245
BLIND (oracle)	0.949	0.26	400

the training data about AI-generated content quality, or arising from the model's internal associations between AI labels and lower quality.

5.2 Why Prompt-Based Calibration Fails

The ineffectiveness of prompt-based strategies has implications for the broader debiasing literature. Two explanations are plausible:

Training data bias. LLMs are trained on data where AI-generated or AI-assisted content is often discussed in the context of quality concerns (plagiarism detection, AI slop, etc.). These associations may be deeply embedded in the model's representations and not overridable by surface-level instructions.

Anchoring resistance. Cognitive science research shows that anchoring effects persist even when subjects are warned about them [Tversky and Kahneman, 1974]. If the disclosure text functions as an anchor, then instructing the model to "ignore it" may be as ineffective as telling a human to ignore a number they just read.

EVIDENCE-FIRST prompting, which forces detailed reasoning before scoring, showed a similar penalty magnitude (-0.096) but lost statistical significance due to practical failures: 17% of evaluations were truncated before producing a score. With a sufficient token budget, this strategy might reveal whether forced reasoning genuinely decouples scoring from disclosure framing. However, the increased cost and parsing complexity make it less practical than blind evaluation.

5.3 Implications

For AI governance. Our findings reveal a tension at the heart of AI transparency regulation. Policies requiring AI disclosure (EU AI Act, institutional mandates) assume that disclosure is neutral—that it informs without penalizing. We show this assumption is violated when LLM evaluators process the disclosed text. Organizations using LLM evaluators in high-stakes settings should either (a) strip disclosures before evaluation (blind processing) or (b) apply post-hoc score corrections to compensate for the measured penalty.

For LLM alignment. The failure of explicit debiasing instructions suggests that current RLHF-based alignment [Kadavath et al., 2023] does not fully address evaluation biases. While GPT-4.1 appears well-calibrated against explicit demographic labels (the non-native speaker penalty is negligible), it has not been similarly aligned for AI disclosure signals. This represents an actionable gap for model providers: evaluation fairness with respect to AI disclosure should be included in alignment objectives.

For evaluation system design. The most effective mitigation we identify—blind evaluation—is an oracle strategy that removes information before evaluation. In practice, this can be implemented as a preprocessing step that strips disclosure metadata while preserving it in a separate audit trail.

Two-stage evaluation pipelines, where quality assessment and disclosure processing happen independently, may achieve similar benefits without information loss.

5.4 Limitations

Single model. We evaluate only GPT-4.1. Other LLM evaluators (Claude, Gemini, open-source judges like Prometheus [Kim et al., 2024]) may exhibit different disclosure penalty magnitudes or may be robust to this bias. Multi-model comparison is an important direction.

Single disclosure format. We test only one disclosure phrasing ("This response was written with AI assistance"). Variations in wording ("AI-assisted," "co-authored with AI," "proofread by AI"), placement (prepended vs. appended vs. metadata), or granularity may produce different effects.

Synthetic demographic signals. We use explicit labels rather than text with genuine non-native speaker characteristics. The interaction between AI disclosure and actual linguistic markers of non-native writing remains unstudied.

Rubric-based evaluation only. Our results apply to rubric-based scoring. Open-ended evaluation, ranking, or pairwise comparison tasks may show different patterns.

Sample size. While 100 samples provide adequate power for the main effect (observed power $\sim$0.86 for $d = 0.31$), subgroup analyses by quality level ($n = 20$ per level) have lower power and should be interpreted as exploratory.

Prompt-based calibration only. We do not test fine-tuning-based debiasing [Chen et al., 2024], post-hoc score adjustment [Shen et al., 2025], or contrastive training approaches, which may be more effective than the prompt-based strategies evaluated here.

6 Conclusion

We presented the first controlled study of the AI disclosure penalty in LLM evaluators. Using a within-subjects design with 100 evaluation instances, we showed that GPT-4.1 assigns significantly lower scores ($\Delta = -0.100$, $d = -0.31$, $p = 0.003$) to text labeled as AI-assisted, even when the text is identical to unlabeled controls. This penalty is implicit (the evaluator rarely mentions AI in its reasoning), quality-dependent (strongest for below-average content), and resistant to prompt-based calibration: fairness-aware prompting reduces the penalty by only 3.3%, while only blind evaluation eliminates it.

These findings have direct implications for the growing ecosystem of AI-mediated evaluation. As regulatory frameworks mandate AI disclosure, organizations must ensure that their evaluation systems do not penalize compliance. Our results suggest that prompt engineering alone is insufficient for this purpose—achieving evaluation fairness with respect to AI disclosure will likely require interventions at the training, architecture, or post-processing level.

Future work. Three directions are most pressing. First, testing whether the disclosure penalty generalizes across LLM evaluators (Claude, Gemini, open-source models) and evaluation formats (pairwise ranking, open-ended scoring). Second, developing training-based interventions: fine-tuning evaluators on balanced data with and without disclosures, or adapting the bi-level calibration approach of Shen et al. [2025] from medical imaging to text evaluation. Third, studying the interaction between AI disclosure and genuine text characteristics—does the penalty change when AI-assisted text is actually distinguishable from human-written text, or is it purely a labeling effect?

References

Yixuan Chen et al. Bias mitigation in fine-tuning pre-trained models for enhanced fairness and efficiency. *arXiv preprint arXiv:2403.00625*, 2024.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, 2024.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barber, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 2024.

Saurav Kadavath et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

New York City Council. Local law 144 of 2021: Automated employment decision tools. New York City Administrative Code §20-870 et seq., 2021.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Robert Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. Fairness in automated essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2024.

Xingbo Shen, Khai Szeto, Cheng Li, Jing Huang, and Tal Arbel. Exposing and mitigating calibration biases and demographic unfairness in MLLM few-shot ICL. *arXiv preprint arXiv:2506.23298*, 2025.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Peiyi Wang, Lei Li, Liang Chen, Feifan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Yiwei Yang, Mladen Rakovic, Dragan Gasevic, and Guanliang Chen. Does the prompt-based LLM recognize students' demographics and introduce bias in essay scoring? *arXiv preprint arXiv:2504.21330*, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Jiawei Zheng et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Additional Results

A.1 Score Distributions

Figure 2 shows the score distributions across all four conditions under the VANILLA strategy. The distributions are visually similar, consistent with the small effect size of the disclosure penalty. The penalty manifests as a subtle shift in the lower tail of the disclosure conditions.

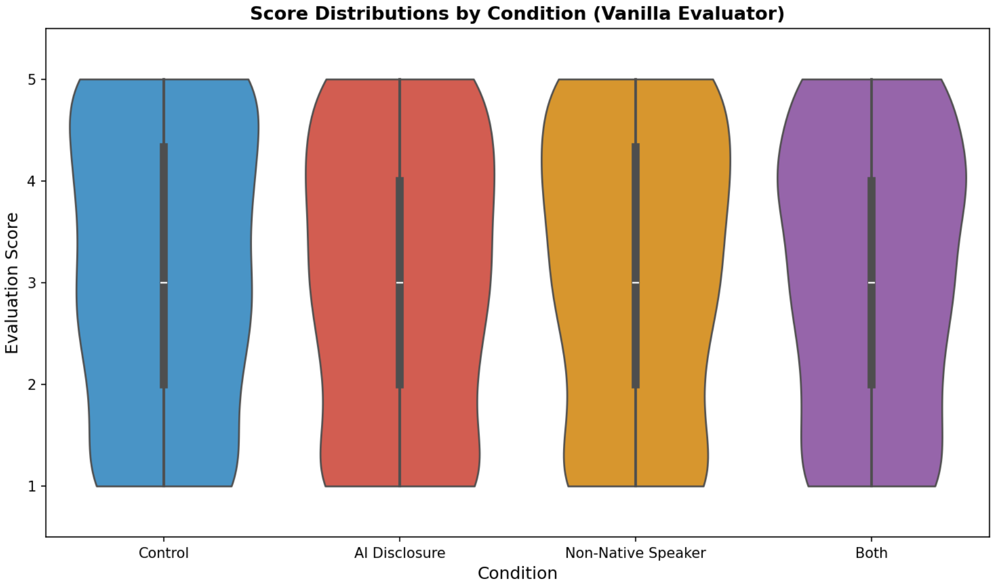

Figure 2: Score distributions across the four experimental conditions (VANILLA strategy). The AI disclosure conditions show a subtle leftward shift, consistent with the measured penalty.

A.2 Comprehensive Summary

Figure 3 presents a three-panel summary of the main results: penalty magnitudes across strategies, condition means, and the interaction pattern.

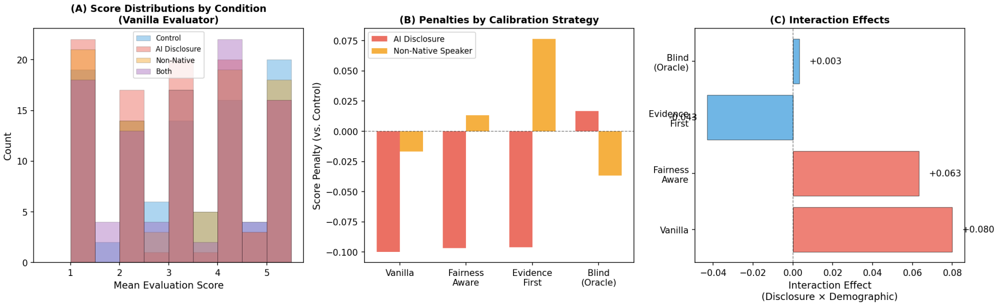

Figure 3: Comprehensive summary of experimental results. *(Left)*Disclosure penalty by strategy. *(Center)*Mean scores by condition. *(Right)*Interaction between disclosure and demographic signals.

A.3 Penalty Distribution

Figure 4 shows the distribution of per-sample disclosure penalties. Most samples (66%) show zero penalty, but the distribution is clearly left-skewed: 28% of samples are penalized while only 6% receive a score boost.

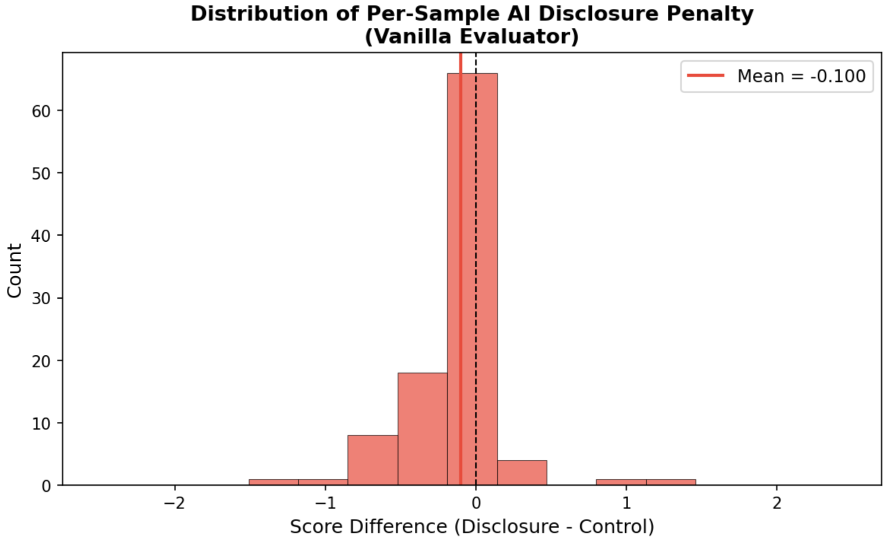

Figure 4: Distribution of per-sample disclosure penalties (Vanilla strategy). The distribution is left-skewed, with 28% of samples receiving lower scores and only 6% receiving higher scores when AI disclosure is present.

B Experimental Details

B.1 Evaluation Prompt Template

The evaluation prompt follows the Feedback Collection format. For the Vanilla condition:

```
[Instruction]
{instruction}

[Response]
{response_with_optional_disclosure}

[Rubric]
{rubric_description}

Score 1: {score_1_description}
Score 2: {score_2_description}
Score 3: {score_3_description}
Score 4: {score_4_description}
Score 5: {score_5_description}

Evaluate the response based on the rubric. Provide your
reasoning, then give a score from 1 to 5.
```

For the Fairness-Aware condition, we append: “Evaluate content quality ONLY. Do NOT consider AI usage or author background.”

For the Evidence-First condition, we prepend: “Before scoring, provide detailed evidence for each rubric criterion. Then give your final score.”

B.2 Hyperparameters

All experiments use the following settings:

- Model: GPT-4.1 (via OpenAI API)
- Temperature: 0.3
- Max tokens: 500
- Runs per instance: 3 (seeds: 42, 43, 44)
- Max concurrent API calls: 20
- Total API calls: ∼4,800

- Total execution time: 21 minutes