Who's the Author? How Explanations Impact User Reliance in AI-Assisted Authorship Attribution

Calvin Bao **Connor Baumler** **Hal Daumé III** **Marine Carpuat**
University of Maryland
{csbao, baumler, hal3, marine}@umd.edu

Abstract

Despite growing interest in explainable NLP, it remains unclear how explanation strategies shape user behavior in tasks like authorship identification, where relevant textual features may be difficult for lay users to pinpoint. To support their analysis of text style, we consider two explanation types: example-based style rewrites and feature-based rationales, generated using a LLM-based pipeline. We measured how explanations impact user behavior in a controlled study (n=95) where participants completed authorship identification tasks with our types of assistance. While no explanation type improved overall task accuracy, fine-grained reliance patterns (Schemmer et al., 2023) revealed that rewrites supported appropriate reliance, whereas presenting both explanation types increased AI overreliance, minimizing participant self-reliance. We find that participants exhibiting better reliance behaviors had focused explanation needs, contrasting with the diffused preferences of those who overrelied on AI, or incorrectly self-relied. These findings highlight the need for adaptive explanation systems that tailor support based on specific user reliance behaviors.

1 Introduction

AI systems have increasingly been used as decision-support tools across domains (Levy et al., 2024; Sel et al., 2024; Echterhoff et al., 2024), despite concerns about overreliance — when people defer decision-making to AI suggestions even when those suggestions are incorrect. Prior work has documented this most clearly in settings involving tasks where AI errors are easily detectable or at least verifiable, as in factual QA (Goyal et al., 2023; Kim et al., 2024). However, we argue that less is known in tasks where decision-making is more interpretive in nature, or where system answers are hard to verify.

This work investigates overreliance and human-AI collaboration in one such setting: authorship attribution. Determining the likely author of a text often depends on subtle stylistic markers which can be difficult for non-experts to pinpoint (Setzu et al., 2024) and differs from how authorship models perform on the task (Faye et al., 2024).

As such, authorship identification offers a case study for understanding how different forms of AI explanations might support or distort user reasoning in tasks where decisions depend on navigating *potentially ambiguous signals or subjective criteria with multiple defensible interpretations*. We consider two LLM-generated explanation types, designed to guide users in making their own assessments of the ambiguous signals inherent in the task: (1) example-based style rewrites, which show a contrastive rewrite of the input to surface changes that might support an alternative decision, inspired by counterfactual explanations, (Lee and Chew, 2023; Cheng et al., 2023; Lee et al., 2024) and (2) feature-based explanations, which highlight and verbalize summary cues that the model uses to faithfully generate its prediction.

We conduct an IRB-approved online user study ($n = 95$) to explore how such explanations impact people's behavior when they are asked to identify the author of a given text given two writing samples from different authors. We designed an experiment with four explanation conditions: (1) AI prediction only, (2) prediction with stylistic rationales, (3) prediction with author-style rewrites, and (4) all support combined. We then evaluated how these strategies affected participant performance, confidence, and alignment with the original prediction, particularly when the AI's prediction was incorrect.

Our work contributes a behavioral lens on human-AI collaboration in authorship decision-making tasks by systematically analyzing how participants respond to system predictions and explanations. As the mere presence of explanations may increase overreliance (Eiband et al., 2019), we characterize a broad range of behavioral patterns

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25312–25330
November 4-9, 2025

(Schemmer et al., 2023), and show how explanations shape these outcomes.

None of the explanation types provided improved overall task accuracy. However, we found that providing only the example-based or the feature-based explanations led to better tradeoffs in correct self-reliance and correct AI-reliance than if users were to use either the dual (example+feature) or no explanations. Additionally, fine-grained analysis suggests that participants with different reliance behaviors have different needs, and different abilities to pinpoint what explanation properties would be useful.

Altogether, results suggest that example-based rewrites are a promising strategy for better appropriate reliance in authorship attribution, and call for adaptive explanation designs tailored to user behavior in explainable NLP to support better reliance behaviors.

2 Background

We motivate our study by discussing prior work in explanations for AI-assisted decision-making and authorship analysis. We focus on how explanations can shape human reliance, and then surface how authorship authorship analysis is an ideal candidate for studying how these explanations impact human decision-making.

2.1 Explanations for AI-Assisted Decision-Making

Explanations can support users in AI-assisted decision-making, by helping users reason how AI systems make predictions. However, prior work has shown that users often defer to AI suggestions despite provided explanations indicating potential errors that those suggestions may not be correct (Bansal et al., 2021). When verifying an explanation is more cognitively demanding than completing the task itself, users may disengage from critical evaluation of explanations and simply accept the suggestion (Vasconcelos et al., 2023). This may stem from users implicitly trusting that explanations signal credibility (Eiband et al., 2019; Lai and Tan, 2019; Sieker et al., 2024), even when local explanations are flawed or when they fail to surface important underlying context that directly impacts task success (Goyal et al., 2024). These findings surface a reliance problem: users must decide when to accept or override AI suggestions, and explanations may or may not help that calibration.

Recent work has explored complementary types of explanations: counterfactual explanations, which make targeted edits to the input to show how the input might look for a model to change its prediction (e.g. (Lee and Chew, 2023; Si et al., 2024)), and example-based explanations (Cheng et al., 2023; Lee et al., 2024), which provide concrete instances of different classes that illustrate the model's decision boundaries.

Another line of research has investigated rationales directly generated by LLMs to explicitly justify given predictions (Wiegreffe et al., 2022; Li et al., 2024; Mishra et al., 2024; Xu et al., 2024). Although LLM-generated rationales have shown improvements in explanation acceptability, these rationales are not consistently reliable (Wiegreffe et al., 2022). To motivate better explanations, there is work explicitly focused on improving contrastive or negative explanations to articulate *why* certain predictions were made instead of alternatives, or providing explanations for non-gold labels to enhance training signals for teaching models how to explain effectively (Wiegreffe and Marasovic, 2021; Wang et al., 2023). We suggest moving beyond focusing on explanation validity alone to identify what properties that make explanations actionable for human decision-making, in order to support both correct self-reliance and AI reliance.

In controlled prediction tasks, Chen et al. (2023) found that example-based explanations helped participants override incorrect AI advice, whereas feature-based explanations increased overreliance. However, their tasks (income prediction and biography classification feature objective predictions which incur relatively low verification cost for people. In contrast, authorship identification requires judging stylistic evidence that is more subjective to verify in practice, leading to higher verification costs. We aim to assess how explanations help users make sense of system predictions, in order to accomplish the task successfully.

2.2 Authorship Analysis

Authorship analysis (such as verification and attribution) is used to power diverse applications and domains (Stamatatos, 2009; Huang et al., 2025), including plagiarisim detection (Quidwai et al., 2023), the detection of machine-generated text (Richburg et al., 2024), forensic analysis (Ainsworth and Juola, 2018), and cultural heritage research (Setzu et al., 2024). Most of this work

focuses on automatic attribution. However, less is known about how human-decision making can be supported in their various contexts.

Given our human–AI lens, the design problem is in *which explanation form best supports human decision-making*. We examine (i) feature-focused rationales and (ii) example-based contrasts. Recent systems commonly realize (i) via stylometry-inspired, LLM-generated rationales (Patel et al., 2023; Hung et al., 2023; Huang et al., 2024; Ramnath et al., 2025; Alshomary et al., 2025). These rationales intend to surface distinctive stylistic aspects, such as lexical or syntactic patterns, which could also serve as explanatory aids to support human judgment in authorship identification tasks. How these explanatory aids could impact human decision-making remains underexplored, which motivates the human study in our work.

3 Study Design

In designing our study, we ask the following research questions:

RQ1. How do different explanation types influence reliance patterns in authorship attribution?

RQ2. What explanation properties do users perceive as most useful for supporting both correct self- and AI- reliance?

We explore these questions by designing a study that explicitly prompts users to try the task by hand, before being exposed to some AI support consisting of explanation-by-example and explanation-by-feature, when they can choose to revise their decision.

Task Overview Figure 1a illustrates the core authorship identification task. Participants are presented with a short **Mystery Post** and two writing samples from two distinct authors (**Candidate A**, **Candidate B**, where one of these is the ground truth author). The task is formulated as a binary classification: which of these authors authored M? The domain is constrained on topic so that participants need to rely on stylistic cues (e.g. lexical choice, syntax, punctuation usage, etc.).

Explanation Interventions Figure 1b illustrates the explanations we propose: **Feature-Based Rationale**: a rationale describing stylistic similarities or differences between the Mystery Post and each candidate's writing. **Example-Based Rewrite**: a rewrite of the Mystery Post in the style of each candidate, offering participants an implicit comparison through example. **Dual Explanation**: both rationale and counterfactual rewrites.

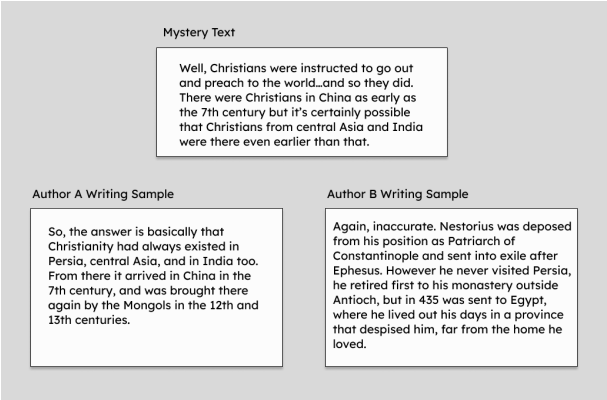
Rationales aim to support user reasoning through explicit comparison, while rewrites aim to provide stylistic transformations as examples, aiming to support user reasoning through implicit comparison through which rewrite shifts the style less.We generate these outputs with LLMs as described in Sections 4.3 and 4.2.

Experimental Setup We adopt a 2×4 mixed design. The within-subject factor is **AI correctness** (correct vs. incorrect prediction), evenly distributed across 10 trials (5 correct, 5 incorrect). The between-subject factor is **explanation type**, randomly assigned to participants, with four conditions: **(1) Prediction-only**: The AI provides a single prediction for the text's author (Author A or Author B). **(2) Dual explanation**: The AI provides a prediction accompanied by explanation-by-feature (stylistic rationales) and explanation-by-example (rewrites illustrating how the mystery post might look if each candidate author had written it). **(3) Explanation-by-feature**: The AI provides a prediction accompanied by the selected explanation-by-feature. **(4) Explanation-by-example**: The AI provides a prediction accompanied by the selected explanation-by-example.

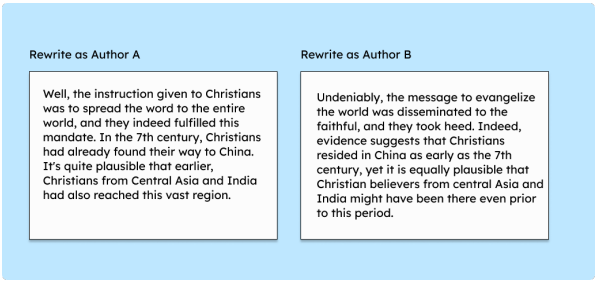
In each trial, participants make an authorship decision in two phases: human-only and AI-assisted. In **Phase 1** (human-only), they choose an author and describe which stylistic cues influenced their decision, with answer choices drawing on interpretable stylistic feature sets (Patel et al., 2023). In **Phase 2** (AI-assisted), participants review the model's binary prediction as well as the provided explanation in their assigned explanation condition, and may revise their choice. After completing all 10 trials, participants complete a paraphrased version of NASA-TLX (Hart and Staveland, 1988) to assess mental workload and perceived task success.

Measures and Analysis We analyze participant behavior in four key aspects:

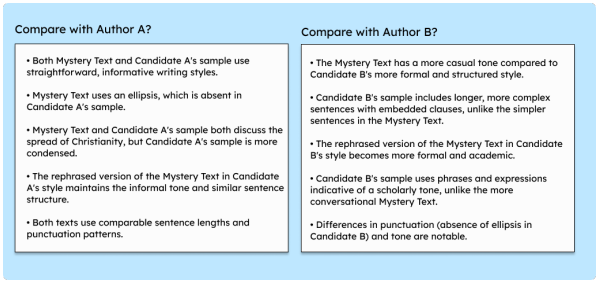
- **Confidence and Accuracy Change**: comparing human-only and AI-assisted confidence and decisions.
- **Reliance Patterns**: categorizing participant decision changes according to AI correctness behavior(Schemmer et al., 2023).



(a) The authorship identification task: the participant is asked which of the two authors wrote the Mystery Post.



(b) The participant receives the prediction, and a counterfactual rewrite as explanation



(c) The participant receives the prediction, and a stylistic rationale as explanation

Figure 1: Two-phase authorship attribution task design. Participants begin with only the Mystery Post and candidate writings (a), and then may revise their decision after seeing AI-generated explanations (b) and/or (c).

- **Cognitive Workload**: using delta NASA-TLX scores across conditions.
- **Cue Use and Explanation Engagement**: analyzing which cues users cite and how explanation needs differ by behavior.

To categorize user reliance patterns, we adopt the *Appropriateness of Reliance* (AoR) framework introduced by Schemmer et al. (2023), which characterizes reliance in terms of two reliance behaviors:

- **Relative AI Reliance (RAIR):** The proportion of cases in which the participant initially made an *incorrect* decision and the AI provided a *correct* prediction, where the participant updated their decision to become correct. Formally:

$$\text{RAIR} = \frac{\text{\# Correct AI Reliance (CAIR)}}{\text{\# trials AI correct}}$$

- **Relative Self-Reliance (RSR):** The proportion of cases in which the participant initially made a *correct* decision and the AI provided an *incorrect* recommendation, where the participant maintained their original correct answer. Formally:

$$\text{RSR} = \frac{\text{\# Correct Self-Reliance (CSR)}}{\text{\# trials AI incorrect}}$$

This analysis allows us to go beyond decision accuracy and evaluate the *cognitive and behavioral effects of explanation format*, with a focus on how different designs shape user reliance.

Participant Recruitment We recruited 95 English-fluent participants on Prolific and compensated them at rates consistent with our local minimum wage. Our survey was implemented in Qualtrics. All participants indicated consent prior to beginning the study after reading a form indicating that their responses would be anonymized and possibly published in aggregate. Our study protocol was approved by our institution's ethics board. We collected basic demographic information (age, education, and self-reported English fluency level) from participants, summarized in Appendix D.

Comparison	Avg. Cosine Similarity
Mystery vs CandA	0.488
Mystery vs CandB	0.475
CandA vs CandB	0.357
RewA vs CandA	0.647
RewA vs CandB	0.433
RewB vs CandA	0.499
RewB vs CandB	0.566

Table 1: Cosine similarities between style embeddings. The top block reports task difficulty (Mystery vs. candidates and candidate–candidate). The bottom block reports rewrite contrastiveness: RewA/RewB vs. intended target and non-target candidates.

4 Task Materials

We construct task materials used in our user study as described in the following sections: (1) **process and select data**, (2) **generate model predictions**, and (3) **derive explanations** (feature-based rationales and counterfactual rewrites).

4.1 Datasets

We use the **PAN 2023 Style Change Detection** dataset[1] which Bevendorff et al. (2023) presented as a challenge task to identify positions within a given multi-author document broken up by when the author switches. Each dataset entry includes multiple passages and a boolean value for each adjacent pair of passages, marking whether or not they were written by different authors. We exploit these markers to construct our (M, A, B) tuples by scanning lines in the hard validation split to find a sequence of three paragraphs: two consecutive paragraphs share the same author, while the other differs. We designate one of the shared posts the Mystery Post and the other two posts the candidate posts. By pulling all three writing samples from the same source document for one authorship trial, we ensure consistent topic content across passages, fulfilling our first desiderata of topical coherence. For evaluation purposes, we sampled 60 tuples randomly from 8.6k potential tuples.

Task Difficulty Check. To verify that trials are comparably hard regardless of which candidate is paired with the Mystery post, we compute cosine similarity between style embeddings (Wegmann et al., 2022) for **(Mystery, Candidate A)**, **(Mystery, Candidate B)**, and **(Candidate A, Candidate B)** (Table 1). The Mystery–Candidate similarities are almost identical (0.488 vs. 0.475), indicating that identifying the author is similarly challenging for both candidates. By contrast, the Candidate–Candidate similarity is lower (0.357), showing that A and B are stylistically distinct. Thus, the task is balanced while still containing discriminative signal, making the core challenge distinguishing the Mystery post from either candidate rather than differentiating A from B.

The task materials are generated by the following models: an Identifier model (subsection 4.2) that produces the authorship prediction, and two explanation generators (subsection 4.3) that produce stylistic feature-focused rationales, and example-based rewrites, which we attach to predictions as experimental conditions for the study.

4.2 How predictions are generated

The **Identifier Model** takes (M, A, B) and returns a label in $\{A, B\}$. We compared open-weight families to select a reliable model for use in our study, leading to a controlled evaluation with three open-weight model families across sizes (Qwen2.5 (Yang et al., 2025), Mistral (Jiang et al., 2023), Llama3 (Grattafiori et al., 2024)), and two verification prompts (LIP, PROMPTAV). We selected PROMPTAV + QWEN2.5-14B based on held-out performance (Appendix B), as it achieved strong correction performance with 0% degradation, ensuring that any gains do not come at the cost of originally correct predictions.

Step 1: Authorship Verification (rationale source). Given (M, A) and (M, B), we prompt the LLM to articulate stylistic similarities or differences, and produce a label or score for each comparison that indicates authorship. Here, we evaluated two prompt designs:

- **LIP** (Huang et al., 2024): prompts an LLM to reason step-by-step about linguistic features (e.g., phrasal verbs, punctuation, sarcasm) and assigns a binary label denoting "same author" or not.
- **PROMPTAV** (Hung et al., 2023): prompts an LLM to focus on linguistic features (e.g. variables such as capitalization, acronyms, and expressions), and report a similarity score (0 to 1) for two texts' writing styles.

[1] https://pan.webis.de/clef23/pan23-web/style-change-detection.html

These justifications serve both as the source for our feature-based rationales, and the input context for one of the judges in Step 2.

Step 2: Authorship Prediction (LLM-as-Judge). We run two parallel LLM-as-judge passes with the same base model, differing only in the evidence they are conditioned on. The first is a rationale-conditioned judge that receives the Step-1 stylistic justifications for (M, A) and (M, B); and the second is a rewrite-conditioned judge that receives counterfactual rewrites of M in A's and B's styles (see subsection 4.3). For sampling, we retain an instance only if *both* judges agree on the predicted author. This agreement filter helps define the final Identifier Model outputs used across conditions.

4.3 How explanations are generated

We describe in more detail how we derived the two types of explanations and use them as the conditioning context for their corresponding judges in the authorship prediction step.

4.3.1 Where feature-based rationale explanations come from

Our **feature-based rationales** are *directly derived* from the stylistic rationales produced during **Step 1: Authorship Verification** in subsection 4.2. The verification prompts (LIP and PROMPTAV) generate short, structured descriptions of stylistic similarities and differences for each pair. We use these justifications directly as feature-based rationales.

4.3.2 Where counterfactual rewrite explanations come from

We generate meaning-preserving rewrites of M in each candidate's style to provide concrete, contrastive examples. Inspired by STYLL (Patel et al., 2024), we implement a streamlined two-step pipeline (Figure 2) using LLAMA3-8b (Grattafiori et al., 2024) as the base rewrite model.

Intended Use of Rewrites. In the user study, we set each of the candidate authors as the "target author" to imitate, and produce one rewrite for each. We expect that the genuine author's rewrite typically requires only minor stylistic shifts, whereas the foil author's rewrite introduces more dramatic changes in style.

Validation of Rewrites. For the rewrites, we assess contrastiveness empirically by comparing each rewritten text (RewA/RewB) to the intended target and non-target candidates (Table 1). We find that each rewrite is closer to its intended target than to the non-target (RewA: **0.647** vs. 0.433; RewB: **0.566** vs. 0.499), validating that the style transfer moves in the intended direction and provides effective rewrite examples to use as anchors.

5 Results and Discussion

Our analysis investigates how different explanation types affect user reliance in AI-assisted authorship attribution (**RQ1**), and what users say or show they need to perform the task better (**RQ2**). We organize our findings into four parts: changes in confidence and accuracy, reliance behavior patterns, perceived workloads of the task, and end with a discussion on what participants rely on to do the task manually, and what unmet needs they perceive to have with respect to the presented explanations.

5.1 Confidence and Decision Accuracy

All AI support inflate confidence. All conditions yielded a significant increase in self-reported confidence after seeing the intervention, regardless of correctness ($p < 0.01$). Confidence gains were largest when stylistic rationales were present (Conditions 2 & 3) (Figure 3), suggesting that rationales increase participant confidence more than rewrites alone. This aligns with prior findings (Vasconcelos et al., 2023; Bansal et al., 2021) that explanations encourage inappropriate reliance on AI.

No accuracy gains; effects depend on explanation format and correctness. When the system was correct, participant accuracy improved modestly across all conditions. However, when the system was incorrect, accuracy dropped in every condition, especially in Condition 2 (Rewrite + Rationale), where the average accuracy decrease was statistically significant ($t = -4.23$, $p = 0.0003$; post-hoc *t*-test on accuracy deltas). This suggests that combining rewrites with rationales increases overreliance. In our authorship attribution setting, explanations are themselves interpretive supports; here, example- vs. feature-based formats did not translate into accuracy gains and the dual format was most fragile under incorrect model predictions.

Manipulation check. To validate that decision changes were indeed driven by explanation support, after each intervention we asked participants why they changed or did not change their choice. Overwhelmingly, participants pointed to the AI prediction, the provided rationales, or the rewrites,

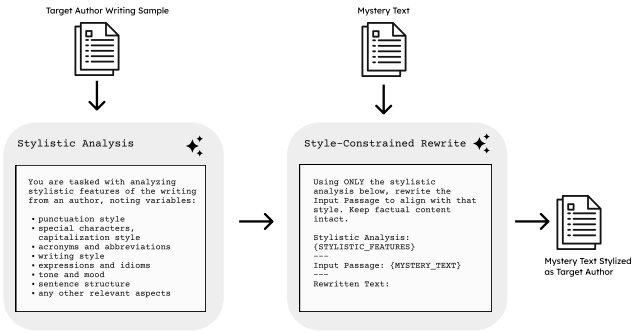


Figure 2: Our Rewrite Model. The first step extracts stylistic variables as target descriptors, as used by Ramnath et al. (2025). The second step then imitates a given target author by rewriting the Mystery Post M using the target descriptors.

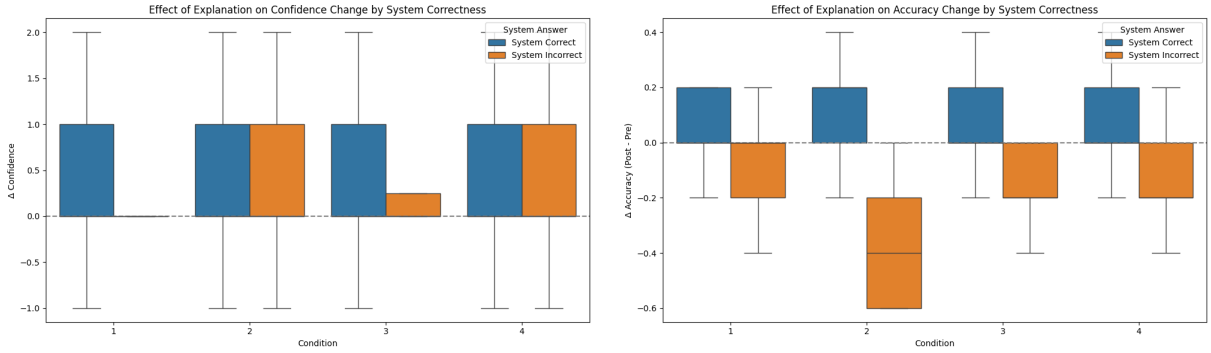


Figure 3: Comparison of explanation effects on (a) accuracy and (b) confidence, each conditioned on whether the system was correct.

with very few citing "No influence" (Table 2). This supports our assumption that the decision shifts are attributable to the experimental interventions.

Summary. Across conditions, explanations reliably increased confidence but did not yield meaningful accuracy improvements. When the model produced incorrect predictions, accuracy decreased, most notably under the combined dual-format explanation. These results suggest that our explanations were persuasive enough to shift decisions but insufficiently able to enable users for task success. We then ask: how are these accuracy decreases impacted by how participants rely on the system?

5.2 Behavioral Patterns of Reliance

To diagnose why confidence rose when exposed to explanations without accuracy gains, we analyze how participants relied on the system using the AoR framework. Table Table 3 summaries these metrics across conditions.

Explanations can support relative AI reliance. All explanation conditions (2-4) show higher relative AI reliance (RAIR) than the prediction-only baseline, indicating that explanations can help participants align appropriately with correct AI outputs. Condition 2 has the highest RAIR (0.157), suggesting it is most effective in improving accuracy when the AI is correct.

Dual-format explanations persuade well. Condition 2's dual combination also drove the lowest relative self-reliance (RSR) (0.322), indicating that users are least likely to resist the AI when it is incorrect. This suggests that participants are persuaded by the dual-format explanation, possibly explained by shortcutting.

Single-format explanations balance AI- and self-reliance. Conditions 3 (rationale-only) and 4 (rewrite-only) achieve a more favorable tradeoff. While not reaching the peak RAIR of Condition 2, they still outperform the prediction-only baseline, and show much higher RSR (especially in condition 4: dual explanations, the highest at **0.542**), indicating stronger resistance to AI errors. This suggests that single-format explanations may offer sufficient support.

What influenced your decision?	Dual	Rationale-only	Rewrite-only
No influence	27	38	46
The authorship prediction	121	108	126
The rewritten versions of the mystery post	91	–	135
The stylistic analysis	134	162	–

Table 2: Self-reported counts toward what influenced participant decision changes during the second phase of an authorship identification trial, broken down by condition. Options not shown to participants in a given condition are not marked.

Condition	Explanation	RAIR ↑	RSR ↑
1	Pred-only	0.108	0.467
2	Dual	**0.157**	0.322
3	Rationale-only	0.142	0.492
4	Rewrite-only	0.125	**0.542**

Table 3: Mean Relative AI Reliance (RAIR) and mean Relative Self-Reliance (RSR) across explanation conditions. Higher RAIR indicates better recovery from errors when the AI is correct; higher RSR indicates better resistance to AI when it is incorrect.

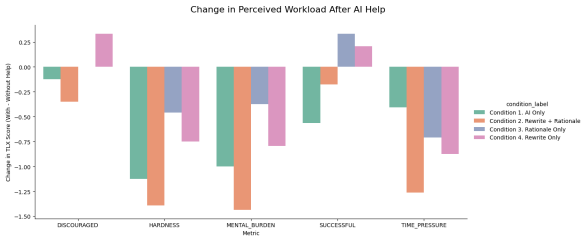


Figure 4: Change in perceived workload dimensions (post–pre) across explanation conditions. Bars reflect mean delta scores. Lower values for MENTAL BURDEN, TIME PRESSURE, SUCCESS indicate better outcomes with AI support. Lower values for DISCOURAGED indicate high discouragement.

5.3 Perception of Workload

To evaluate how explanations influenced various aspects of perceived workload, we analyzed deltas from the NASA-TLX survey across five dimensions: mental burden, time pressure, task success, discouragement, and hardness. As shown in Figure 4, we observed no increase in reported *mental burden* across any condition, suggesting that no explanation was considered cognitively demanding.

Dual explanations yielded the sharpest decline in perceived mental burden and time pressure, possibly reflecting a false sense of trust when they are provided, aligning with its lower RSR (increased overreliance). Single explanations yielded mild increases in discouragement and slightly lower perceived task success, suggesting that they may have made participants more aware of ambiguity, and more reliant on themselves to verify task success.

These patterns align with our behavioral observations. Dual explanations may reduce perceived effort while encouraging inappropriate trust, while single explanations led users to think critically.

5.4 Stylistic Cue Usage and Explanation Unmet Needs

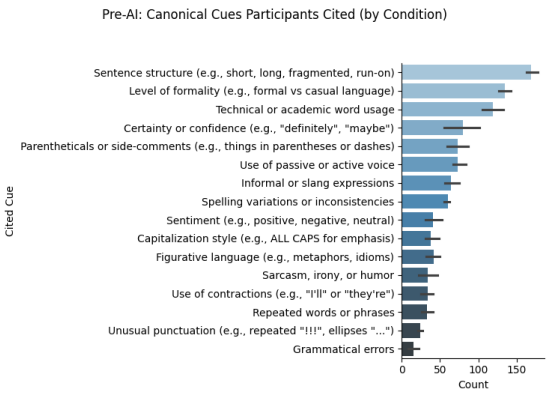


Figure 5: Stylistic cues participants identified as useful for attribution. Sentence structure, formality, and technical or academic word usage dominated cue selection.

We analyze two questions: which stylistic cues participants actually use for the task, and which additional explanations they say would have helped them with the task.

Across all conditions, the most frequently cited cues included sentence structure, formality level, and technical or academic word usage. These preferences suggest that participants relied mostly on surface-level and structural cues that are both visually salient and easier to contrast between authors.

We categorize participants by their dominant reliance behavior. Within each category, we compute each participant's aggregate preference rank of requested explanations across trials, then report

Explanation Type	CAIR	CSR	IAIR	ISR
1. Highlighted keywords	✗*	✓	✓	✓
2. Similar example post	✓*	✓	✓	✗
3. Writing style summary	✗	✓	✗*	✓*
4. Heatmap over post	✓	✗	✓	✗
5. No additional explanation	✗	✗	✓	✗*

Table 4: Preferences for explanation types across dominant reliance patterns. ✓indicates relative preference (lower average rank); ✗indicates relative dispreference. A superscript *flags the extreme within a column: ✗*= least-preferred; ✓*= most-preferred.

group-level preferences. We report whether the unseen explanation is relatively preferred (✓) vs relatively dispreferred (✗) in Table 4.

When people follow the AI: correct (CAIR) vs. incorrect (IAIR). Both groups who went with the AI's answer liked seeing a similar example post. But CAIR participants leaned toward heatmaps over the posts and were least positive about highlighted keywords. IAIR participants asked for many different kinds of help: they liked highlighted keywords, similar example posts, heatmaps, and even no additional explanation. In this context, participants displaying CAIR had focused explanation needs, while participants displaying IAIR felt that many different explanations would have helped them in this task.

When people rely on themselves: correct (CSR) vs. incorrect (ISR). Both participant groups exhibiting correct self-reliance (CSR) and incorrect self-reliance (ISR) preferred ✓writing style summaries (most preferred by ISR), and highlighted keywords, but diverged on similar example posts (CSR would prefer, ISR would not), and strikingly, ISR felt like they did not need additional explanations to do the task, though they may have needed explanations that persuasively explains why a system prediction is correct.

Future Explanation Design Participants who exhibit more appropriate reliance behaviors (CAIR and CSR) appear to have clearer and more targeted explanation needs. In contrast, participants with less desirable reliance behaviors (IAIR and ISR) request many available explanations, indicating a general desire for support without a clear sense of what would be useful. From a design perspective, this highlights the importance of adaptive explanation strategies: effective systems should reinforce the successful behaviors of appropriately calibrated users while identifying how to better scaffold decision-making for users who struggle.

Implications for Authorship Attribution Because people have no outside knowledge to check decisions against, they rely heavily on how the AI frames the decision. We see this when participant confidence increased across the board after AI input, even when the overall accuracy did not. The way those signals are packaged as explanations has implications for appropriate reliance. Our results highlight what makes authorship attribution uniquely challenging for human-in-the-loop. Participants are required to reason about subtle stylistic cues that are often hard to pinpoint, and even verify, as we empirically see that many failed to recover from initial mistakes (incorrect self-reliance).

6 Conclusion

We investigate how explanation-by-feature (rationale), explanation-by-example (rewrite), and their combination affect human decision-making in an authorship attribution context. Our results show that single-format explanations hit the right balance in introducing positive friction in their decision-making process, while combining the two in the dual explanation may have led to unintentionally reducing perceived effort by leading them to short circuit. These findings highlight a challenge in explainable AI: more explanations, even if they do not increase mental burden, can inflate confidence while degrading decision accuracy. Our explanation preference analyses suggests a path forward – people who exhibit appropriate reliance behaviors tend to prefer more targeted, low-friction explanations. Additionally, the single-format explanations we test promote better self-reliance. Future work should explore more adaptive explanations that promote correct self-reliance and help users recognize when AI advice warrants revision.

Limitations

Our study has several limitations. First, while authorship attribution is a cognitively demanding task, our study setup is low-stakes for participants, potentially limiting the generalizability of our findings to real-world decision-making scenarios with higher consequences. Second, we do not provide a clear ground truth or rationale to users, which may have influenced participant engagement differently than tasks with more objective criteria and verifiable outcomes. While this ambiguity reflects real-world challenges in interpreting AI decisions, it also complicates the interpretation of accuracy and reliance metrics. Our participant pool, though diverse in demographics, consists of online crowdworkers, which introduces potential selection biases. These users may differ from domain experts or casual end-users in how they perceive AI support or seek explanations. Additionally, our explanation strategies are limited to two types—rationales and rewrites—generated via prompting large language models. Future work could explore a broader design space of explanations.

Potential Risks

Our work carries potential risks related to overreliance on AI systems in difficult-to-verify decision-making tasks. While our goal is to understand and mitigate inappropriate reliance on AI, the explanation strategies we study could be misused to increase user compliance regardless of accuracy or quality. If deployed uncritically, such explanations may lead users to defer to AI outputs even when those outputs are incorrect or biased. To mitigate these risks, we advocate for adaptive explanation designs that support calibrated reliance and encourage user reflection.

Acknowledgements

We would like to thank the members of the CLIP lab at the University of Maryland for their constructive feedback and support. Specifically, we would like to thank Kem Nguyen-Le, Dayeon Ki, and Yimin Xiao for their valuable feedback in our study, and Navita Goyal and Fumeng Yang for their input on prior discussions of this project.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Janet Ainsworth and Patrick Juola. 2018. Who wrote this: Modern forensic authorship analysis as a model for valid forensic science. *Washington University Law Review*, 96:1159–1190.

Milad Alshomary, Narutatsu Ri, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, and Kathleen McKeown. 2025. Latent space interpretation for stylistic analysis and explainable authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1124–1135, Abu Dhabi, UAE. Association for Computational Linguistics.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2023. Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 459–481, Cham. Springer Nature Switzerland.

Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2023. Relic: Investigating large language model responses using self-consistency. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms.

Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA. Association for Computing Machinery.

Géraud Faye, Benjamin Icard, Morgane Casanova, Julien Chanson, François Maine, François Bancilhon, Guillaume Gadek, Guillaume Gravier, and Paul Égré. 2024. Exposing propaganda: an analysis of stylistic cues comparing human annotations and machine classification. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 62–72, Malta. Association for Computational Linguistics.

Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 155–180, New York, NY, USA. Association for Computing Machinery.

Navita Goyal, Eleftheria Briakou, Amanda Liu, Connor Baumler, Claire Bonial, Jeffrey Micher, Clare Voss, Marine Carpuat, and Hal Daumé III. 2023. What else do I need to know? the effect of background information on users' reliance on QA systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3330, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, et al. 2024. The llama 3 herd of models.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *SIGKDD Explor. Newsl.*, 26(2):21–43.

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA. Association for Computing Machinery.

Min Hun Lee and Chong Jun Chew. 2023. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2518–2531, New York, NY, USA. Association for Computing Machinery.

Sharon Levy, William D. Adler, Tahilin Sanchez Karver, Mark Dredze, and Michelle R. Kaufman. 2024. Gender bias in decision-making with large language models: A study of relationship conflicts.

Tang Li, Mengmeng Ma, and Xi Peng. 2024. Beyond accuracy: Ensuring correct predictions with correct rationales. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing large language models as rationalizers of knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8117–8139, Bangkok, Thailand. Association for Computational Linguistics.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2024. Low-resource authorship style transfer: Can non-famous authors be imitated?

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Ali Quidwai, Chunhui Li, and Parijat Dube. 2023. Beyond black box AI generated plagiarism detection: From sentence to document level. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 727–735, Toronto, Canada. Association for Computational Linguistics.

Sahana Ramnath, Kartik Pandey, Elizabeth Boschee, and Xiang Ren. 2025. CAVE: Controllable authorship verification explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8939–8961, Albuquerque, New Mexico. Association for Computational Linguistics.

Aquia Richburg, Calvin Bao, and Marine Carpuat. 2024. Automatic authorship analysis in human-AI collaborative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1845–1855, Torino, Italia. ELRA and ICCL.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 410–422, New York, NY, USA. Association for Computing Machinery.

Bilgehan Sel, Priya Shanmugasundaram, Mohammad Kachuee, Kun Zhou, Ruoxi Jia, and Ming Jin. 2024. Skin-in-the-game: Decision making via multi-stakeholder alignment in llms.

Mattia Setzu, Silvia Corbara, Anna Monreale, Alejandro Moreo, and Fabrizio Sebastiani. 2024. Explainable authorship identification in cultural heritage applications. *J. Comput. Cult. Herit.*, 17(3).

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474, Mexico City, Mexico. Association for Computational Linguistics.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The illusion of competence: Evaluating the effect of explanations on users' mental models of visual question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19459–19475, Miami, Florida, USA. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(1):538–556.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

A Prompting Details

In this section, we include the full prompt templates used for authorship identification in Step 2. Prompt2A and Prompt2B correspond to the baseline and intervention conditions, respectively, as described in the main text.

Prompt 2A: Vanilla Authorship Identification

Task: You are an expert in stylistic analysis. Below you are given:

- A base text (the "Mystery Text").
- Two candidate author samples.
- A breakdown of stylistic elements assessing whether or not the base text and one candidate author text was written by the same author ("MYSTERY POST WRITTEN BY CANDIDATE [A,B]?")

Your task is to assess the comparisons of the stylistic features of the Mystery Text against Candidate A and Candidate B.
Then, in a final statement, identify which candidate is most likely the author of the Mystery Text based on the presented stylistic evidence as well as your analysis.

Mystery Text: {MYSTERY TEXT}

Candidate 1: {CANDIDATE 1 TEXT}

Candidate 1 Verification Rationale: {Stylistic Analysis between Candidate 1 and Mystery text}

Candidate 2 Verification Rationale: {Stylistic Analysis between Candidate 2 and Mystery text}

Candidate 2: {CANDIDATE 2 TEXT}

Figure 6: Prompt2A: Baseline LLM Judge Prompt. Given the mystery text, two candidate texts, and verification rationales, the model selects the most likely author.

B Identifier Model Evaluation

Table 5 summarizes the results of our evaluation comparing authorship accuracy, correction rate, degradation rate, and neutral rates across multiple models and verification prompting methods (LIP and PromptAV) with their counterfactual variants. For completeness, we include also a baseline **Direct** attribution system, which directly prompts an LLM to make an authorship attribution decision (with and without rewrites).

Prompt 2B: Counterfactual Authorship Identification

Task: You are an expert in stylistic analysis. Below you are given:

- A base text (the "Mystery Text").
- Two candidate author samples.
- Two stylistically rewritten versions of the Mystery Text: one rewritten in Candidate A's style and one rewritten in Candidate B's style.

Your task is to assess the stylistic features of the Mystery Text, its rewritten versions, and the candidate texts. First, compare the Mystery Text with Candidate A's sample and its rewritten version. Then, compare the Mystery Text with Candidate B's sample and its rewritten version. Based on your analysis, provide a final statement indicating which candidate is most likely the author of the Mystery Text.

Mystery Text: {MYSTERY TEXT} **Candidate 1**: {CANDIDATE 1 TEXT} **Mystery Text Rewritten in Candidate 1's Style**: {Mystery Text Rewritten with CANDIDATE 1 Style descriptors} **Mystery Text Rewritten in Candidate 2's Style**: {Mystery Text Rewritten with CANDIDATE 2 Style descriptors}

Figure 7: Prompt2B: Proposed LLM Judge Prompt. In addition to the inputs in Prompt2A, the model is given counterfactual rewrites of the mystery text based on each candidate author's style.

Counterfactual Accuracy The PromptAV method consistently shows accuracy improvements across all models evaluated. The largest accuracy improvement was observed in the Qwen2.5-7B PromptAV setting.

The setting with PromptAV supplied with the

Model	Method	Original Accuracy	Counterfactual Accuracy	Correction Rate	Degradation Rate
Qwen2.5-7B	Lip	0.585	0.623	11.3%	7.5%
Qwen2.5-7B	PromptAV	0.472	0.642	28.3%	11.3%
Qwen2.5-7B	Direct	0.600	0.567	11.67%	15.0%
Qwen2.5-14B	Lip	0.517	0.600	13.3%	5.0%
Qwen2.5-14B	PromptAV	0.567	0.667	10.0%	0.0%
Qwen2.5-14B	Direct	0.916	0.916	3.3%	3.3%
Qwen2.5-32B	Lip	0.767	0.817	8.3%	3.3%
Qwen2.5-32B	PromptAV	0.783	0.883	13.3%	3.3%
Qwen2.5-32B	Direct	0.915	0.983	6.77%	0%
Mistral-7B	Lip	0.617	0.717	21.7%	11.7%
Mistral-7B	PromptAV	0.517	0.567	16.7%	11.7%
Mistral-7B	Direct	0.728	0.847	22.03%	10.17%
Mistral-24B	Lip	0.723	0.660	4.3%	10.6%
Mistral-24B	PromptAV	0.745	0.787	10.6%	6.4%
Mistral-24B	Direct	0.9	0.767	5.0%	18.33%
Llama-3.1-8B	Lip	0.667	0.617	5.0%	10.0%
Llama-3.1-8B	PromptAV	0.683	0.717	8.3%	5.0%
Llama-3.1-8B	Direct	0.896	0.827	6.9%	13.79%

Table 5: Evaluation for authorship attribution accuracy, counterfactual accuracy, and correction rates. For convenience, we bold the better between the "Original Accuracy" and "Counterfactual Accuracy". "Correction Rate" refers to the percentage of predictions initially incorrect or unsure corrected by the rewrite intervention. "Degradation Rate" refers to the percentage of predictions initially correct made incorrect by the rewrite intervention. We focus on the "explain-then-attribute" pipelines to extract the model's stylistic evidence and test how targeted rewrites influence decisions. While a direct classifier generally attains higher accuracy, it lacks the extractions of explicit evidence / rationales, which limits our user-facing experiments. We therefore tradeoff accuracy for manipulability and measurable correction rates under controlled edits.

counterfactual rewrites raises accuracy score above each original baseline, with steepest lift coming from Qwen-2.5-7B, where PromptAV boosts counterfactual accuracy by roughly 17 points (0.472 $\rightarrow$ 0.642). The counterfactual configuration for PromptAV outperforms its respective configuration in LIP in 5 of the 7 models—the exceptions being Mistral-7B (0.717 vs. 0.567). **Direct** generally achieves higher raw accuracy than LIP or PromptAV, and it is of interest that in this domain and of the LLMs we test, the direct approach leads to better task accuracy than feature-extraction-first pipelines.

Correction and Degradation Rate. The correction rate measures the proportion of instances where the intervention successfully corrected previously incorrect predictions. The degradation rate, conversely, measures the proportion of instances where the intervention makes incorrect the initially correct prediction. In Table 5, there are two exceptions (Mistral-24B and Llama-8B) where the correction rate is lower than the degradation rate.

Given the combined insights from accuracy improvements, correction and degradation rates, we select **PromptAV with Qwen2.5-14B** as our final Predictor Model due to it having the least degradation rate (0%). Counterfactual accuracy alone can hide trade-offs: a model might boost its score simply by flipping as many correct predictions to wrong ones as it fixes, leaving users with an unpredictable tool. Our rationale for selecting the model with the minimum degradation rate guarantees that no originally correct decisions are sacrificed for new gains. We also treat **Direct** as a performance reference rather than a model whose outputs we use in our human study, as we lack extractable natural language features.

Although this evaluation shows that counterfactual rewrites can reasonably enhance performance on the attribution task, it remains unclear whether such rich, whole-text explanations help or hinder human decision-making. To investigate this, we next run a controlled user study to examine whether participants benefit from, and at what point they may overrely on, explanations based on counterfactual rewrites and rationales.

C Survey Task Instructions

Figure 8 shows the screen participants see when they begin the study. Figure 9 shows **Phase 1** of a

given authorship trial. Figure 10 shows **Phase 2** of a given authorship trial.

D Participant Demographics

We asked only for age, education level, and English fluency level, summarized in Table 6.

Demographic	Count
Age	
18–24 years old	11
25–34 years old	35
35–44 years old	19
45–54 years old	13
55–64 years old	11
65+ years old	6
Education	
High school diploma or GED	12
Some college, no degree	5
Associates or technical degree	2
Bachelor's degree	41
Graduate or professional degree	35
English Fluency	
Very well	91
Well	4

Table 6: Participant demographic summary (age, education, and English fluency; $n = 95$).

E Distribution of Behavioral Patterns.

To contextualize these aggregate scores, we show the distribution of the observed reliance behaviors. Figure 11 shows the full distribution of reliance behaviors. We note that:

- **Incorrect AI Reliance (IAIR)** (**C** $\rightarrow$ **I** | **I**) was most frequent in Condition 2.
- **Correct Self-Reliance (CSR)** (**C** $\rightarrow$ **C** | **I**) was highest in Conditions 1 and 4.
- **Correct AI Reliance (CAIR)** (**I** $\rightarrow$ **C** | **C**) was low across all conditions, suggesting limited support for correction, even with explanations.
- **Incorrect Self-Reliance (ISR)** (**I** $\rightarrow$ **I** | **C**) was low across all conditions, suggesting limited support for correction, even with explanations.

Thank you for agreeing to take part in this study!

Your job is to help us understand how people attribute short text passages to different authors—and how AI assistance (predictions, rewrites, or stylistic highlights) impacts those judgments.

In each question, you will see the following:
1. A **Mystery Post** -- a short excerpt with unknown authorship
2. Two **Candidate Posts**, each written by a different author

Your task is to decide which candidate author most likely wrote the mystery post **based only on stylistic features.** These are things like writing style, tone, phrasing, sentence structure, and word choice — **not** based on the topic or content itself. For example, one author might tend to write short, direct sentences, while another uses longer, more descriptive ones. Or one author might use a very formal, precise tone, while another could use a casual, conversational tone.

You will then be asked to rate your confidence and answer a few follow-up questions. Depending on the version of the study, you will also see additional AI assistance before making your choice.

There are *10 trials overall*, and the whole session should take about 15 minutes. Please read each screen carefully, answer to the best of your ability, and try not to go back once you've submitted your response.

All responses are anonymous and will only be used for research.

Figure 8: Introduction to the study task.

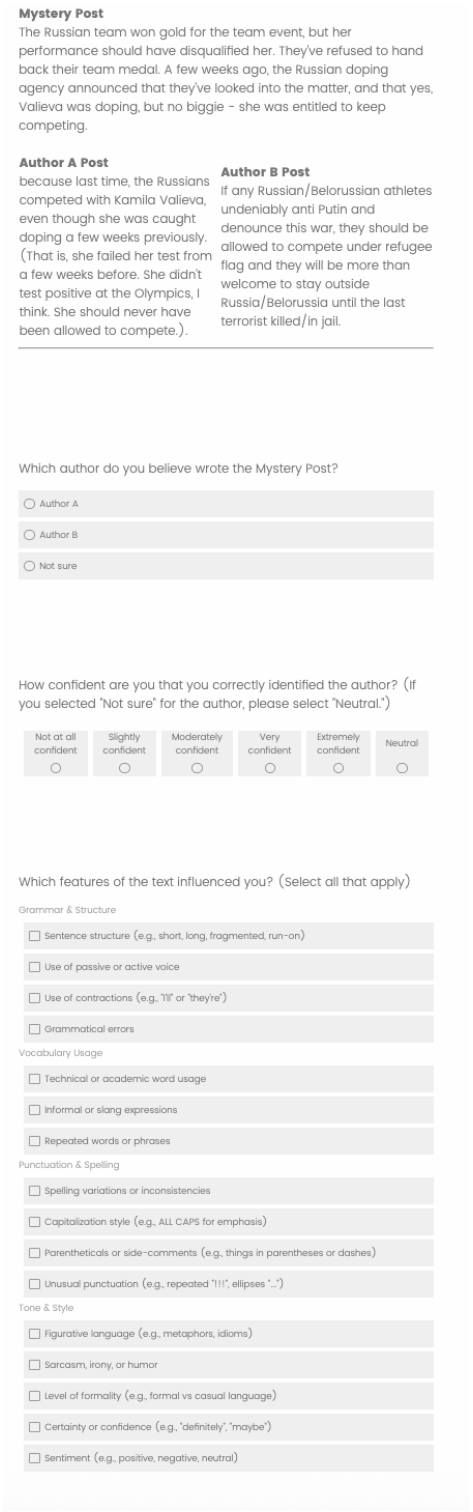


Figure 9: Task instructions for **Phase 1**.

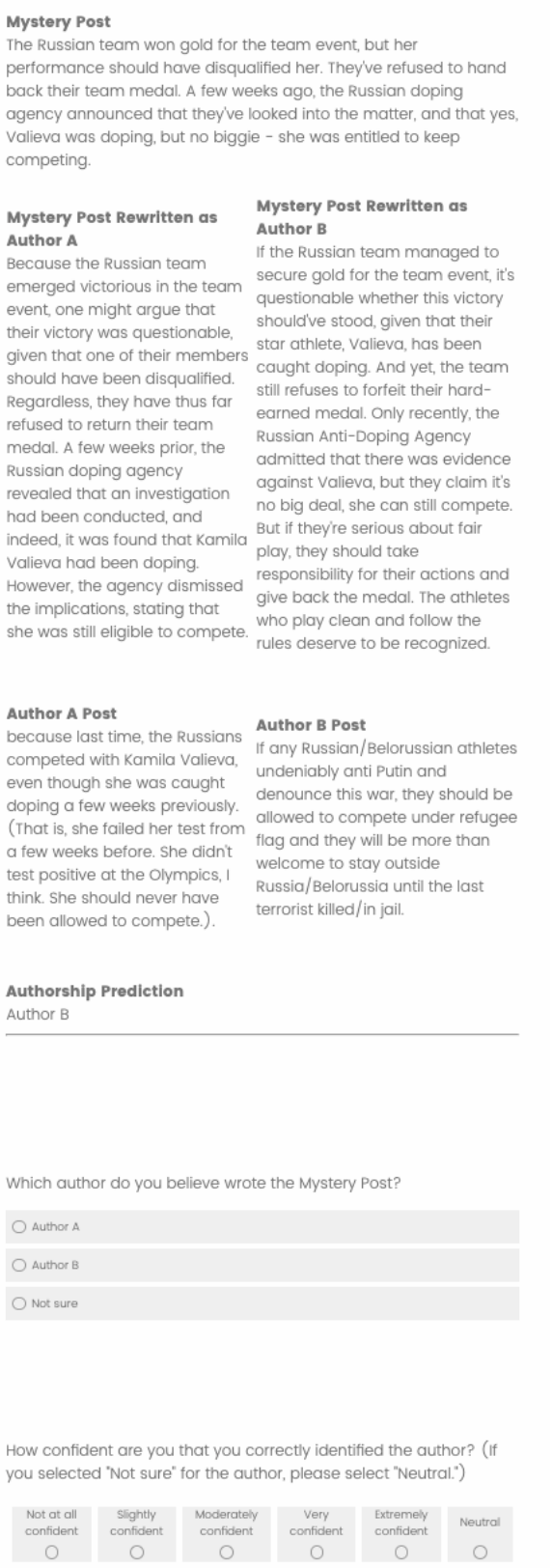


Figure 10: Task instructions for **Phase 2**.

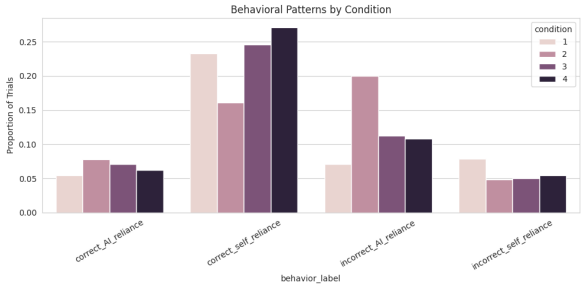


Figure 11: Distribution of key reliance behaviors by condition. Notably, Condition 2 (Rewrite + Rationale) shows the highest incorrect AI reliance and lowest self-reliance, indicating that our participants shift toward trusting the AI.