

When LLMs Say No: Diagnosing and Mitigating Conservative Bias in Authorship Verification

Nicholas Wlezien

University of Houston

ncwlezie@cougarnet.uh.edu

Rakesh M. Verma

University of Houston

rmverma2@central.uh.edu

Abstract

Large language models (LLMs) have increasingly been used for zero-shot authorship verification, where a model must determine whether two texts share an author. We show that evaluated zero-shot prompting methods exhibit a conservative bias, often over-predicting different authorship. This behavior is especially pronounced when the texts of the same author differ in topic or content. Through qualitative and quantitative analyses, we find evidence that LLMs confuse differences in content with differences in authorship. To test whether this failure can be reduced, we introduce lightweight prompt interventions that instruct the model to focus on style rather than content. Across four datasets and five open-weight instruction-tuned LLMs from three model families, the best intervention improves same-author accuracy by 13.1 points while also increasing macro-F1 by 3.1 points on average across all 20 experimental settings. Our findings show that zero-shot authorship verification is sensitive to the separation of content and style and that small framing changes can meaningfully change the direction of LLM errors. Code is available at <https://github.com/galactixx/when-llms-say-no>.

1 Introduction

Authorship verification (AV) asks whether two texts were written by the same author. Given a pair (T_1, T_2) , an AV system predicts $y \in \{\text{YES}, \text{NO}\}$, where YES denotes same authorship. AV builds on stylometry, authorship attribution, and cross-domain verification (Juola, 2008; Stamatatos, 2009; Koppel et al., 2011; Kestemont et al., 2020), and is relevant to forensic linguistics, authorship disputes, and academic-integrity settings. LLMs now make AV possible in zero-shot settings, where a user can prompt a model to compare two texts and return a confidence score or YES/NO decision. Existing AV prompts instruct models to reason across stylistic features including punctuation, capitalization,

idioms, tone, and sentence structure (Hung et al., 2023; Ramnath et al., 2025). These methods are beneficial because they don't require labeled examples or training and can generate high-quality explanations. However, prompted LLMs may mistake content differences for authorship differences. When two same-author texts differ in topic, character names, or genre, the model may treat those differences as evidence for NO. This is especially problematic for cross-domain AV, where the target phenomenon is authorial style rather than topical similarity. Prior work has long noted that topic and style are difficult to disentangle (Seroussi et al., 2014; Sari et al., 2018; Kestemont et al., 2020) and zero-shot LLM settings may amplify this confound. We study whether zero-shot LLM AV prompts exhibit a conservative NO bias and whether this bias is driven by content conflation. We ask:

1. Do existing zero-shot AV prompts overestimate NO?
2. Are same-author pairs with larger semantic distance more likely to be misclassified as NO?
3. Can small prompt interventions reduce false-negatives by pushing models to focus more on stylistic over topic evidence?

We introduce two prompt interventions. The first is a *framing* instruction that tells the model not to focus on topic, and the second is a *contrastive reweighting* instruction that instructs the model to separate low-level stylometric evidence from content-sensitive evidence. Each intervention is added as a separate instruction to recent prompting methods PromptAV (Hung et al., 2023) and Prompt-CAVE (Ramnath et al., 2025), and evaluated as variants. Our contributions are:

- We identify conservative NO prediction as a central failure mode for zero-shot LLM authorship verification (§5).

- We diagnose this behavior through qualitative inspection of model outputs and a semantic-distance analysis of same-author pairs (§6).
- We introduce simple prompt interventions that tell the model to prioritize low-level stylistometric over content-sensitive evidence (§7).
- We show that the best-performing interventions reduce false-negative errors, where the strongest intervention improves YES accuracy by 13.1 percentage points on average across all 20 experimental settings, while increasing macro-F1 by 3.1 points (§8–9).

2 Related Work

Authorship verification and stylometry. Authorship verification (AV) decides whether two documents share an author. Classical work relies on stylometric evidence including character and word n -grams, function words, punctuation, and syntactic patterns (Juola, 2008; Stamatatos, 2009; Koppel et al., 2011). In addition, cross-domain benchmarks show that systems must generalize across topics and genres rather than exploit topical overlap (Bevendorff et al., 2020; Kestemont et al., 2020). Recent work extends this to cross-genre authorship analysis. Ma et al. (2025) introduce CrossNews, a benchmark for authorship verification and attribution with Article–Article, Article–Tweet, and Tweet–Tweet settings. AV results are also sensitive to dataset construction and evaluation protocol (Työ et al., 2022). Our work extends this to zero-shot LLMs that generate natural-language explanations without training.

Topic, style, and content confounding. A central difficulty in authorship analysis is separating author style from topic. Features that appear predictive of authorship may instead capture what an author writes about, which topics appear in the dataset, or which genres are associated with particular authors (Seroussi et al., 2014; Sari et al., 2018). Similarly, Wegmann et al. (2022) show that content control is important for learning style representations independent of topic. The topic-confusion setting makes this explicit by changing author–topic associations between training and testing (Altakrori et al., 2021). This concern is especially relevant when using LLMs for AV where a model may remark on differences in subject matter, named entities, or setting and treat them as evidence for different authorship. We study this

failure mode by asking whether false-negative errors increase as the semantic distance within pairs becomes larger.

LLM prompting for authorship analysis. Prompting offers a zero-shot route to authorship analysis. In this setting the model is instructed to compare two texts and return a confidence score or YES/NO decision, along with an explanation. PromptAV asks an LLM to compare stylistic variables and produce a same-author confidence score (Hung et al., 2023). Prompt-CAVE extends this setup with structured feature-level explanations and a final YES/NO decision (Ramnath et al., 2025). Similarly, Linguistically Informed Prompting (LIP) provides explicit linguistic guidance to improve LLM performance on authorship verification and attribution while preserving natural-language explanations (Huang et al., 2024). Recent benchmark work also evaluates LLMs for authorship identification more broadly. AIDBench studies same-author and one-to-many candidate identification across emails, blogs, reviews, articles, and research papers (Wen et al., 2024). These approaches show that LLMs can perform zero-shot authorship judgments and that linguistic prompting can improve aggregate performance. However, they do not analyze whether prompted LLMs make directional errors or whether those errors concentrate on semantically distant same-author pairs. To do this, we measure semantic distance using `all-mpnet-base-v2` embeddings¹ as a proxy for content divergence (Reimers and Gurevych, 2019; Song et al., 2020).

LLM scoring and style transfer approaches. One-shot style transfer (OSST) estimates authorship similarity through LLM log probabilities, targeting the separation of style and content (Miralles-González et al., 2025). Our work is complementary. Instead of changing the scoring mechanism, we analyze how zero-shot prompted LLMs reason toward false-negative decisions. Taken together, prior work establishes that stylometric evidence can support AV, that LLM prompts can use linguistic cues, and that separating style from content is a challenge. We build on this work by diagnosing false-negative bias, linking it to semantic divergence, and testing prompt interventions that prioritize stylistic evidence over content-sensitive cues.

¹`all-mpnet-base-v2`

Dataset	Total	YES	NO
Blog-Auth	1,000	500	500
Fanfic	1,000	500	500
IMDb62	1,000	503	497
CrossNews	1,000	500	500

Table 1: Test-split statistics. YES/NO = same-author/different-author pairs.

3 The Diagnostic Target: Conservative NO Bias

3.1 Task Definition and Class-Directional Error

Each example is a text pair (T_1, T_2) with a binary gold label $y \in \{\text{YES}, \text{NO}\}$. The model returns a confidence score $s \in [0, 1]$, a YES/NO label, or a structured explanation. For score-based prompts, we map scores to YES only when $s > 0.5$ and to NO otherwise. We define conservative bias as a tendency for the model to under-predict YES and over-predict NO. On balanced datasets, high NO accuracy and low YES accuracy can yield deceptively moderate aggregate scores, so we report overall accuracy, macro-F1, and class accuracy. The central question is whether zero-shot prompt AV systems fail directionally, specifically accepting different-author pairs while rejecting same-author pairs whose content differs.

3.2 Content Conflation Hypothesis

Our central hypothesis is that false-negative errors arise when the model conflates semantic distance with authorship difference. Let $d(T_1, T_2)$ denote semantic distance between the two texts. Among gold-YES pairs, we expect the probability of a false-negative error to increase with semantic distance:

$$P(\hat{y} = \text{NO} \mid y = \text{YES}, d(T_1, T_2)) \text{ increases with } d(T_1, T_2). \quad (1)$$

We first test this by fitting a logistic regression with `statsmodels` (Seabold and Perktold, 2010) per baseline prompt (PROMPTAV & PROMPT-CAVE) that models false-negative errors as a function of semantic distance, and dataset and model fixed-effects. Further, we group gold-YES pairs into semantic-distance quintiles to understand if the false-negative rate increases along with semantic-distance.

4 Experimental Protocol

4.1 Datasets

We evaluate on four datasets. These include the three CAVE benchmark datasets (IMDb62, BLOG-AUTH, FANFIC) covering movie reviews, personal blogs, and fan fiction (Seroussi et al., 2014; Schler et al., 2006; Kestemont et al., 2020; Ramnath et al., 2025), and a sampled subset of CROSSNEWS (Ma et al., 2025). When creating the CROSSNEWS test split we sample from the Article–Article subset, drawing a stratified random sample of 1,000 pairs balanced across classes. Each example contains two pieces of text and a YES/NO label. Basic test-split statistics can be seen in Table 1.

4.2 Models and Inference

We evaluate five open-weight instruction-tuned LLMs from three model families. These include Gemma 3 12B-IT/27B-IT (Gemma Team, 2025), Qwen3-14B/32B (Qwen Team, 2025), and Mistral-Small-3.2-24B-Instruct (Mistral AI, 2025). We abbreviate them as G3-12B, G3-27B, Q3-14B, Q3-32B, and Mist-24B. All models use greedy decoding with temperature 0 and a 4096-token budget, served with batched vLLM inference (Kwon et al., 2023) on a single H100 GPU. We use no system prompt, each example is rendered as a user-only chat message, and Qwen3 models use non-thinking mode. We purposely do not tune thresholds for score-based methods as tuning thresholds would require labeled data which is not available in a zero-shot environment.

4.3 Prompt Families and Interventions

For every experimental setting which consists of a model–dataset pair, we run nine zero-shot methods, consisting of three baselines and six intervention variants. Because our goal is to diagnose directional errors made in zero-shot LLM prompting, we restrict the baselines to LLM-based AV systems and leave the testing of non-LLM systems to future work.

Chain-of-thought score prompt. The COT baseline asks the model to reason step by step and provide a confidence score from 0 to 1 indicating the likelihood that the two texts were written by the same author.

PromptAV. The PROMPTAV baseline extends the score prompt by listing stylistic variables relevant to AV, including punctuation style, capital-

Model	Mth	Blog-Auth		Fanfic		IMDb62		Cross-News	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
G3-12B	COT	0.59	0.55	0.53	0.42	0.69	0.67	0.73	0.71
	PAV	0.54	0.44	0.51	0.37	0.61	0.55	0.65	0.62
	CAVE	0.53	0.42	0.50	0.34	0.54	0.42	0.61	0.56
G3-27B	COT	0.59	0.56	0.56	0.50	0.67	0.66	0.72	0.71
	PAV	0.58	0.55	0.57	0.52	0.67	0.66	0.72	0.72
	CAVE	0.62	0.59	0.54	0.44	0.72	0.70	0.78	0.78
Mist-24B	COT	0.65	0.64	0.56	0.54	0.69	0.69	0.68	0.66
	PAV	0.64	0.63	0.60	0.60	0.69	0.69	0.77	0.76
	CAVE	0.59	0.60	0.57	0.60	0.64	0.67	0.77	0.79
Q3-14B	COT	0.56	0.48	0.52	0.44	0.59	0.53	0.59	0.51
	PAV	0.56	0.49	0.52	0.44	0.60	0.55	0.63	0.59
	CAVE	0.47	0.41	0.39	0.36	0.46	0.42	0.51	0.50
Q3-32B	COT	0.60	0.57	0.55	0.47	0.67	0.66	0.71	0.69
	PAV	0.60	0.58	0.58	0.58	0.69	0.69	0.75	0.75
	CAVE	0.59	0.59	0.59	0.61	0.71	0.74	0.81	0.84

Table 2: Baseline Acc and macro-F1. PAV=PROMPTAV, CAVE=PROMPT-CAVE. Bold = best per model.

ization, acronyms and abbreviations, writing style, expressions and idioms, tone and mood, sentence structure, and other relevant aspects (Hung et al., 2023). The output is parsed for a decimal score, and a fixed threshold maps the score to YES or NO.

Prompt-CAVE. The PROMPT-CAVE baseline asks for a structured response in a JSON format with evidence per feature, a final score, and an output label (Ramnath et al., 2025). We extract the output label to arrive at a final answer and further analyze consistency between feature judgments, final score, and the label.

Framing. The framing instruction tells the model to disregard topic and prioritize style:

Do not focus on what the texts are about.
Focus on recurring surface-form habits
that could remain stable across topics.

Contrastive reweighting. The contrastive instruction tells the model to separate low-level stylistometric evidence from content-sensitive evidence:

When comparing the texts, explicitly
separate low-level stylistometric evidence
from content-sensitive evidence, and
give more weight to low-level stylistometric
evidence.

Variants. We apply these interventions to both PROMPTAV and PROMPT-CAVE prompts. This results in three additional prompt variants for each family consisting of framing-only, contrastive-only, and framed-contrastive. This design isolates the

Model	Mth	Blog-Auth		Fanfic		IMDb62		Cross-News	
		YES _A	NO _A	YES _A	NO _A	YES _A	NO _A	YES _A	NO _A
G3-12B	COT	0.28	0.91	0.09	0.97	0.48	0.90	0.47	0.98
	PAV	0.13	0.95	0.04	0.98	0.25	0.98	0.35	0.96
	CAVE	0.08	0.98	0.00	1.00	0.09	0.99	0.25	0.98
G3-27B	COT	0.33	0.85	0.22	0.89	0.48	0.87	0.51	0.93
	PAV	0.29	0.88	0.24	0.89	0.47	0.87	0.62	0.81
	CAVE	0.36	0.88	0.12	0.96	0.47	0.96	0.66	0.91
Mist-24B	COT	0.53	0.77	0.33	0.79	0.55	0.83	0.42	0.94
	PAV	0.49	0.78	0.52	0.68	0.63	0.75	0.70	0.83
	CAVE	0.39	0.79	0.61	0.54	0.63	0.64	0.82	0.73
Q3-14B	COT	0.18	0.93	0.14	0.91	0.23	0.95	0.20	0.98
	PAV	0.18	0.94	0.13	0.91	0.25	0.96	0.30	0.97
	CAVE	0.07	0.86	0.05	0.72	0.07	0.86	0.12	0.90
Q3-32B	COT	0.32	0.88	0.17	0.93	0.48	0.86	0.45	0.97
	PAV	0.35	0.86	0.46	0.69	0.57	0.81	0.74	0.76
	CAVE	0.37	0.82	0.57	0.61	0.69	0.74	0.79	0.83

Table 3: Class-conditional accuracy. YES_A/NO_A = per-class accuracy. Bold = best YES_A and NO_A per model.

effects of each instruction in addition to their combination. Full prompt templates for all nine methods are included in Appendix G (Table 22). For the camera-ready experiments, we standardized prompt templates to remove differences in formatting and intervention placement. All results reported in this paper use these standardized templates.

4.4 Parsing and Metrics

For prompt variants whose answer is dependent on a numeric score (PROMPTAV & COT) the last decimal is extracted and used to derive the final score. Scores greater than 0.5 are mapped to YES, otherwise NO. The outputs where a score cannot be parsed are counted as incorrect but are not assigned to either the YES or NO class. For PROMPT-CAVE variants we extract the first JSON object and parse it with strict JSON, Python literal parsing, and `json-repair` (Baccianella, 2025), then read output as YES, NO, or MAYBE and final score as the confidence score. Outputs of MAYBE are counted as incorrect predictions regardless of the gold label and are not assigned to either class. Invalid CAVE outputs, those that could not be parsed, are handled the same way. Sensitivity results for PROMPT-CAVE with MAYBE outputs excluded and for PROMPTAV with invalid outputs excluded are both reported in Appendix E.

5 Existing Prompts Already Lean Toward NO

Table 2 reports baseline accuracy and macro-F1 across the three prompt families and five models.

Setting	Gold / Pred	Baseline rationale excerpt
CROSSNEWS Q3-14B COT row 278	YES / NO	“While the style and tone are similar, the subject matter is entirely different, and the context and content suggest different reporting angles and events.”
BLOG-AUTH G3-27B PROMPTAV row 201	YES / NO	“While both texts use standard capitalization and a limited number of abbreviations, these similarities are not strong enough to suggest a common author, especially given the overwhelming differences in other areas. The subject matter is also drastically different.”
IMDB62 Q3-14B PROMPT-CAVE row 412	YES / NO	“Both texts are well-written and coherent, but they differ in subject matter and emotional tone. No other significant stylistic differences are observed.”

Table 4: Representative baseline rationales for false-negative predictions on gold-YES pairs. In each case the model acknowledges some stylistic similarities, then predicts different authorship on subject matter, context, or tone. Additional examples in Tables 9 and 17.

Baseline	Coef.	Odds ratio	p -value
PROMPTAV	0.635	$1.89\times$	2.34×10^{-62}
PROMPT-CAVE	0.921	$2.51\times$	6.30×10^{-82}

Table 5: Pooled logistic regression of false-negative errors on gold-YES pairs with valid YES/NO outputs as a function of semantic distance, controlling for dataset and model. Standard errors are clustered by text pair. Higher semantic distance significantly increases the odds of false-negative errors among true same-author pairs. Odds ratios are per one standard-deviation increase in semantic distance.

No single baseline dominates across all models. COT is the best across all datasets for G3-12B, PROMPTAV is the strongest overall for Q3-14B and remains competitive for Mist-24B, while PROMPT-CAVE is strongest in most model–dataset cells for G3-27B and Q3-32B. Notably, as the number of parameters increases, PROMPT-CAVE becomes much better relative to the other baselines.

Aggregate metrics understate the problem and Table 3 makes the directional failure clear. Each of the baseline prompts generally shows high NO accuracy alongside much lower YES accuracy. However, this effect is not as pronounced in several PROMPT-CAVE cells for Q3-32B and reverses in two Mist-24B cells. In 65% (39/60) of the cells, NO accuracy is above 0.80 while YES accuracy is below 0.50, meaning models behave as conservative verifiers that frequently reject same-author pairs.

The asymmetry is strongest on FANFIC and BLOG-AUTH and especially for smaller models. G3-12B PROMPT-CAVE achieves near-perfect NO accuracy on FANFIC while recovering very few YES examples, while Q3-14B PROMPTAV has YES ac-

curacy no higher than 0.18 on BLOG-AUTH. At the aggregate level this asymmetry appears across all three baselines, although a few individual model–dataset cells show exceptions.

6 Diagnosing the Cause: Content Conflation in Model Rationales

The class-bias pattern shows a directional failure, but not why same-author pairs are rejected. We therefore inspect baseline rationales on gold-YES pairs misclassified as NO. Across datasets and models, we observe rationales that treat content, genre or topic as reasons for different authorship, even when surface-level evidence points the other way.

Table 4 gives three representative false-negative rationales. In each, the model names stylistic similarities but still predicts NO after emphasizing content-related differences such as subject matter or context. These are examples of content conflation, where relevant stylometric evidence is present but overridden by topical dissimilarity.

If content conflation drives false-negative errors, semantically distant same-author pairs should be misclassified more often. Table 5 reports pooled logistic regressions on gold-YES pairs, with standard errors clustered by text pair to account for the same pair being evaluated across multiple models. Each text is embedded using `all-mpnet-base-v2` which encodes the first 384 tokens of each text. Higher semantic distance significantly increases false-negative odds for both baselines, $1.89\times$ for PROMPTAV ($p = 2.34 \times 10^{-62}$) and $2.51\times$ for PROMPT-CAVE ($p = 6.30 \times 10^{-82}$).

Together, the rationales and regressions support the mechanism that models treat topic as a proxy

Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
BASE	0.58	0.54	0.29	0.88	0.55	0.50	0.28	0.83	0.65	0.63	0.43	0.87	0.70	0.69	0.54	0.87
+FR	0.59	0.56	0.35	0.84	0.56	0.54	0.41	0.71	0.66	0.65	0.52	0.80	0.72	0.72	0.64	0.81
+CO	0.58	0.53	0.28	0.88	0.55	0.51	0.36	0.74	0.65	0.63	0.46	0.85	0.70	0.69	0.56	0.84
+BOTH	0.58	0.55	0.34	0.82	0.55	0.55	0.50	0.60	0.66	0.65	0.56	0.76	0.73	0.72	0.68	0.77

Table 6: PROMPTAV interventions, averaged across five models. Acc = accuracy, F1 = macro-F1, YES_A/NO_A = per-class accuracy. Bold = best per dataset column. Per-model results in Table 15.

Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
BASE	0.56	0.52	0.25	0.87	0.52	0.47	0.27	0.77	0.61	0.59	0.39	0.84	0.70	0.69	0.53	0.87
+FR	0.58	0.56	0.34	0.82	0.53	0.51	0.37	0.68	0.64	0.65	0.52	0.76	0.72	0.74	0.66	0.78
+CO	0.54	0.49	0.19	0.89	0.51	0.44	0.19	0.82	0.59	0.56	0.30	0.89	0.69	0.68	0.49	0.88
+BOTH	0.56	0.54	0.27	0.85	0.52	0.50	0.34	0.69	0.63	0.63	0.42	0.83	0.70	0.72	0.60	0.80

Table 7: PROMPT-CAVE interventions, averaged across five models. Acc = accuracy, F1 = macro-F1, YES_A/NO_A = per-class accuracy. Bold = best per dataset column. Per-model results in Table 16.

for authorship difference. Admittedly, the regression is not causal proof that semantic distance alone determines the prediction, as we utilize semantic distance as a proxy for content divergence. However, its value is diagnostic and shows that false-negative errors concentrate where the hypothesized confound should be strongest. In addition, the qualitative outputs show that models often verbalize exactly this confound. This motivates the interventions and subsequent prompt variants we introduce and test.

7 Prompt Reframing Reduces False-Negative Errors

Tables 6 and 7, which report results averaged across the five models, show that the framing-based variants generally shift the error profile toward more YES predictions. However, the effect depends on both the intervention and prompt family, as the PROMPT-CAVE contrastive-only variant decreases YES accuracy on average. The intended effect is not simply to raise overall accuracy, but to correct the directional imbalance identified above by recovering more true same-author pairs without completely sacrificing different-author discrimination.

For PROMPTAV, combining framing with contrastive reweighting produces the largest YES accuracy gains, at the cost of lower NO accuracy, although it is only a partial tradeoff as macro-F1 increases on average. For PROMPT-CAVE, framing alone performs best overall, as YES accuracy increases 13 points on IMDB62 while macro-F1 also increases by 6 points. Overall, framing is the most effective instruction. It is the best standalone

variant for PROMPT-CAVE, forms part of the best-performing PROMPTAV variant, and remains competitive as a standalone PROMPTAV intervention. However, the differences in performance across families suggest that the most effective intervention depends on the base prompt format.

8 Statistical Significance of Intervention Gains

Table 8 tests each intervention against its family baseline across all 20 experimental settings using Holm-adjusted one-sided McNemar tests for YES accuracy and Holm-adjusted one-sided paired permutation tests implemented with `SciPy` (Virtanen et al., 2020) for macro-F1.

For PROMPTAV, framing plus contrastive reweighting improves YES accuracy in 18 of 20 settings (+13.1 points; 13 Holm-significant at $p < .05$). The framing variant alone improves YES accuracy in 19 of 20 settings (+9.3 points; 12 Holm-significant at $p < .05$). For PROMPT-CAVE, framing alone improves in 19 of 20 settings (+11.2 points; 16 Holm-significant at $p < .05$). The contrastive variant for PROMPT-CAVE improves only two settings for YES accuracy. For the variants that improve YES accuracy on average, the macro-F1 gains indicate partial rebalancing rather than simply reversing the bias. The PROMPT-CAVE contrastive variant is the exception, reducing both YES accuracy and macro-F1.

Family	Variant	YES Accuracy			Macro-F1		
		Mean Δ	Impr.	McNemar sig. $p < .05/.01$	Mean Δ	Impr.	Perm. sig. $p < .05/.01$
PROMPTAV	Framing	+9.3	19/20	12/20 / 12/20	+2.9	14/20	7/20 / 6/20
	Contrastive	+2.8	13/20	6/20 / 5/20	+0.1	8/20	3/20 / 3/20
	Framing + Contrastive	+ 13.1	18/20	13/20 / 13/20	+ 3.1	12/20	7/20 / 6/20
PROMPT-CAVE	Framing	+ 11.2	19/20	16/20 / 16/20	+ 4.8	15/20	11/20 / 11/20
	Contrastive	-6.7	2/20	0/20 / 0/20	-2.5	3/20	0/20 / 0/20
	Framing + Contrastive	+4.8	14/20	11/20 / 10/20	+2.7	13/20	7/20 / 7/20

Table 8: Statistical evidence for the prompt interventions. Mean Δ values are percentage-point changes averaged over 20 model–dataset settings relative to the corresponding family baseline (PROMPTAV variants vs. PROMPTAV; PROMPT-CAVE variants vs. PROMPT-CAVE). “Improved” = number of settings with positive delta. For YES accuracy, significance counts report Holm-adjusted one-sided McNemar tests for improvement over the corresponding baseline. For Macro-F1, significance counts report Holm-adjusted one-sided paired permutation tests for improvement over the corresponding baseline. Bold = best per family.

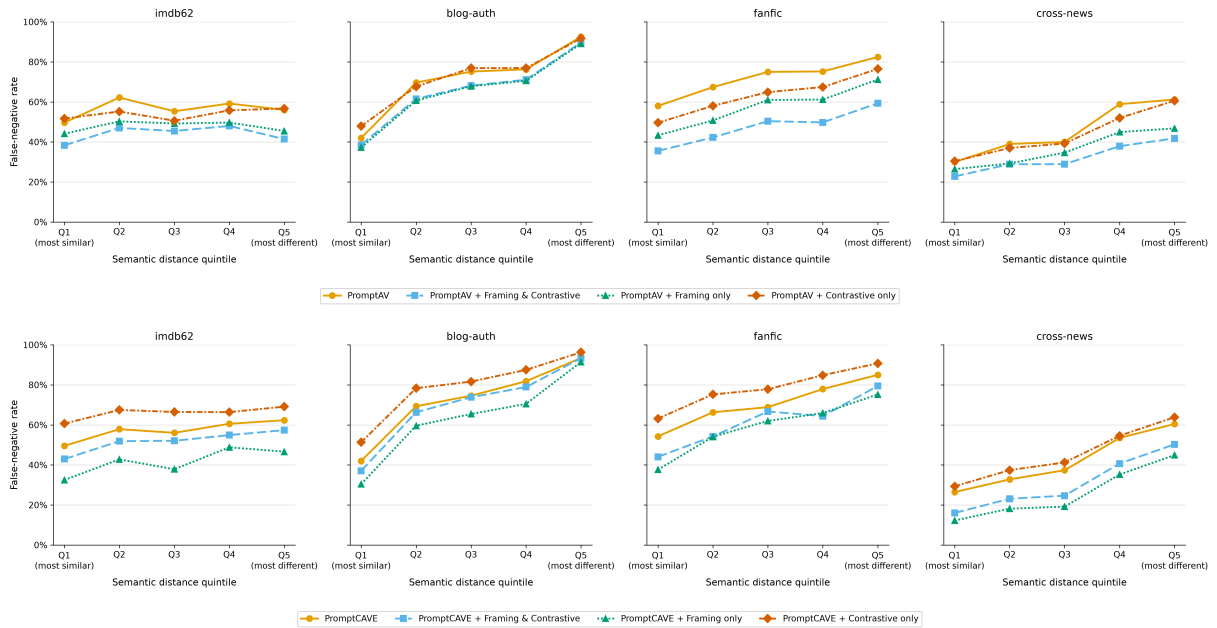


Figure 1: False-negative rate on gold-YES pairs with valid YES/NO outputs by semantic-distance quintile, averaged across the five evaluated models and shown separately by dataset. Q1 contains the most semantically similar same-author pairs and Q5 the most semantically distant. Top: PROMPTAV variants. Bottom: PROMPT-CAVE variants. The plots show that semantic distance is associated with more false-negative errors, especially on BLOG-AUTH, FANFIC, and CROSSNEWS, while the best-performing interventions, PROMPTAV with framing plus contrastive reweighting and PROMPT-CAVE with framing, reduce the false-negative rate approximately equally across quintiles.

9 Semantic Distance, Intervention Gains, and Qualitative Evidence

The results in §5 show that models tend to over-predict NO. We next ask where those errors concentrate. We compute semantic distance with all-mpnet-base-v2 encoding the first 384 tokens of each text and rank gold-YES pairs into quintiles from most similar (Q1) to most distant (Q5). Figure 1 plots false-negative rates by quintile for PROMPTAV and PROMPT-CAVE. If content conflation drives the bias, errors should rise from Q1 to Q5.

False-negative rates generally rise from Q1 to Q5, especially on BLOG-AUTH and FANFIC. The best intervention curves are lower than the baselines approximately equally across all quintiles. Framing plus contrastive reweighting helps PROMPTAV, while framing alone helps PROMPT-CAVE. BLOG-AUTH remains hardest, suggesting that register and discourse-type variation is a stubborn conflation cue. While false-negative rates decrease, visual inspection suggests that the slopes do not flatten, indicating that the association does not disappear. Per-model curves are in Appendix F.

Setting	Outcome	Baseline rationale excerpt	Intervention rationale excerpt
FANFIC Mist-24B PROMPTAV +FR row 71	Gold: YES Base: NO Interv.: YES	“... the texts share some stylistic similarities (punctuation, capitalization, sentence structure) but differ significantly in tone, mood, setting, and idiomatic usage. The genres are also distinct (fantasy vs. crime/mystery), which is a strong clue that they were written by different authors.”	“... the texts share many surface-form habits... The differences in tone and specific idioms are minor and could be attributed to different scenes or characters rather than different authors.”
BLOG-AUTH Q3-32B PROMPTAV +BOTH row 649	Gold: YES Base: NO Interv.: YES	“There are some stylistic similarities, but key differences in content, tone, and expression suggest different authors.”	“Given the moderate level of strictness, and the strong low-level stylometric similarities, we can be moderately confident that the two texts were written by the same author.”
CROSSNEWS Mist-24B PROMPT-CAVE +FR row 533	Gold: YES Base: NO Interv.: YES	“Text 1 focuses on a specific individual (Cressida Dick) and a TV show, while Text 2 discusses historical and documentary analysis. The subjects and perspectives are unrelated.”	“Both texts exhibit a similar tendency toward digressive commentary and layered arguments.”

Table 9: Representative false-negative corrections. In all three cases, the baseline emphasizes differences in content, genre, or tone and predicts different authorship, while the intervention redirects attention toward recurring stylistic habits and recovers the gold same-author label. Additional examples are in Table 17.

Table 9 shows individual predictions where the baseline labels NO and the intervention flips to YES. The baseline attends to content while the intervention discounts those cues and focuses on stylistic habits, resulting in a correct outcome.

10 Reliability, Trade-offs, and Broader Implications

The intervention gains should be interpreted alongside output reliability. The relevant reliability measures include consistency metrics and MAYBE and invalid rates for each PROMPT-CAVE variant, as well as invalid rates for PROMPTAV variants. Full reliability tables are in Appendix A. Further, the PROMPT-CAVE results are robust to excluding MAYBE outputs, while PROMPTAV results are robust to excluding invalid outputs (Appendix E). Threshold adjustment can recover some YES examples, but does not remove the underlying conflation. We verify this in Appendix B through a manual audit of 600 near-threshold gold-YES PROMPTAV outputs, 92.7% of which contained NO-leaning rationales (Table 14).

Tuning the threshold (τ^*) would require labeled examples which are not available in a zero-shot setting. Moreover, τ^* varies substantially across datasets. The proposed interventions change what evidence the model attends to, not merely how a score is mapped to a label.

11 Conclusion

Overall, zero-shot LLM AV prompts can be strongly conservative, over-predicting different-author labels. Lightweight task-framing interventions can partially reduce the failure, but the best intervention depends on prompt family and carries a YES/NO trade-off. The broader implication is that prompt-based evaluators can substitute an easier surface cue for the intended judgment. Evaluations should report class-directional errors and test whether errors concentrate in predictable sub-regions of the input space.

12 Limitations

This study is limited to zero-shot prompting with a fixed set of open-weight instruction-tuned models, datasets, and decoding settings. The trade-off between improved YES recovery and reduced NO accuracy means the best prompt may depend on application costs. Semantic distance is only a proxy for content difference, so we treat it as a mechanism diagnostic rather than causal proof. Structured-output reliability also remains a limitation for CAVE-style prompting. Future work should test whether these findings extend to closed-source models and other instruction-tuning objectives.

Ethical Considerations

Authorship verification can be used in high-stakes settings such as academic integrity investigations,

legal disputes, and forensic analysis. LLM-based AV systems should not be used as sole evidence for authorship decisions. This work exposes and partially mitigates one failure mode, but does not eliminate risks of misclassification, dataset bias, or misuse. We use only publicly released datasets, models, and software libraries, cite their original creators, and follow their stated licenses and terms of use. Our use is limited to research evaluation and is consistent with their intended research use and access conditions. We do not redistribute restricted datasets or model weights.

References

- Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. [The topic confusion task: A novel evaluation scenario for authorship attribution](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Stefano Baccianella. 2025. [JSON Repair: A python module to repair invalid JSON, commonly used to parse the output of LLMs](#). Software.
- Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. [Shared tasks on authorship analysis at PAN 2020](#). In *Advances in Information Retrieval*, pages 508–516. Springer International Publishing.
- Gemma Team. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- 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.
- 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.
- Patrick Juola. 2008. [Authorship attribution](#). *Foundations and Trends in Information Retrieval*, 1(3):233–334.
- Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. [Overview of the cross-domain authorship verification task at PAN 2020](#). In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*, Thessaloniki, Greece. CEUR-WS.org.
- Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. [Authorship attribution in the wild](#). *Language Resources and Evaluation*, 45(1):83–94.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, Koblenz, Germany. Association for Computing Machinery.
- Marcus Ma, Duong Minh Le, Junmo Kang, Yao Dou, John Cadigan, Dayne Freitag, Alan Ritter, and Wei Xu. 2025. [CrossNews: A cross-genre authorship verification and attribution benchmark](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24777–24785, Philadelphia, Pennsylvania, USA.
- Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2025. [LLM one-shot style transfer for authorship attribution and verification](#). *arXiv preprint arXiv:2510.13302*.
- Mistral AI. 2025. [Mistral small 3.2](#). Model documentation for Mistral Small 3.2.
- Qwen Team. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. [Topic or style? exploring the most useful features for authorship attribution](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. [Effects of age and gender on blogging](#). In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium, AAAI Technical Report SS-06-03*, pages 199–205. AAAI Press.

- Skipper Seabold and Josef Perktold. 2010. [statsmodels: Econometric and statistical modeling with python](#). *SciPy 2010*.
- Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. [Authorship attribution with topic models](#). *Computational Linguistics*, 40(2):269–310.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [MPNet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.
- Efstathios Stamatatos. 2009. [A survey of modern authorship attribution methods](#). *Journal of the American Society for Information Science and Technology*, 60(3):538–556.
- Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. [On the state of the art in authorship attribution and authorship verification](#). *arXiv preprint arXiv:2209.06869*.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental algorithms for scientific computing in python](#). *Nature Methods*, 17(3):261–272.
- 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.
- Zichen Wen, Dadi Guo, and Huishuai Zhang. 2024. [AIDBench: A benchmark for evaluating the authorship identification capability of large language models](#). *arXiv preprint arXiv:2411.13226*.

Dataset	Variant	CS-1	CS-2	Combined
Blog-Auth	BASE	91.7	91.4	90.9
	+FR	90.4	90.3	89.4
	+CO	90.8	87.0	86.3
	+BOTH	90.3	88.1	87.6
Fanfic	BASE	87.7	86.9	84.9
	+FR	85.5	87.5	84.1
	+CO	85.1	82.9	80.1
	+BOTH	82.7	82.6	80.3
IMDb62	BASE	87.4	87.8	85.2
	+FR	85.0	86.7	83.7
	+CO	85.0	83.8	80.6
	+BOTH	82.6	82.7	80.3
CrossNews	BASE	87.3	85.9	83.6
	+FR	87.1	87.1	84.6
	+CO	84.1	82.9	79.8
	+BOTH	83.8	82.9	80.6

Table 10: Output consistency of PROMPT-CAVE variants, averaged across five models. CS-1 = agreement between the final score and final output label; CS-2 = agreement between the feature-level judgments and the final output label within the same run; Combined = both criteria met. Bold = best per dataset. Interventions generally reduce output consistency relative to the baseline.

A Output Reliability Tables

Tables 10–12 report the output reliability diagnostics summarized in Section 10.

B Threshold Sensitivity and Calibration

Table 13 reports the effect of optimizing the decision threshold on PROMPTAV baseline outputs across all four datasets. For each model–dataset combination, we find the threshold τ^* that maximises accuracy on the test set and compare performance at our fixed decision threshold ($\tau = 0.5$) against performance at τ^* .

Thresholding and calibration. The threshold analysis clarifies the conservative-bias result. Several baselines retain useful ranking information, and lowering the score threshold can improve test-set accuracy and recover additional same-author pairs. However, this is a post-hoc calibration change as it changes how an already generated score is mapped to a YES/NO, rather than changing the model’s original prompted judgment. Further, Table 13 shows that the optimal threshold varies substantially by dataset, ranging from 0.324 on BLOG-AUTH to 0.464 on CROSSNEWS.

Manual audit of near-threshold false-negative rationales. To test whether near-threshold errors

Dataset	Variant	MAYBE rate	Invalid rate
Blog-Auth	BASE	7.3%	0.0%
	+FR	7.8%	0.0%
	+CO	6.9%	0.0%
	+BOTH	7.6%	0.0%
Fanfic	BASE	9.5%	0.0%
	+FR	9.4%	0.0%
	+CO	7.9%	0.0%
	+BOTH	11.9%	0.0%
IMDb62	BASE	8.9%	0.0%
	+FR	10.4%	0.0%
	+CO	8.4%	0.0%
	+BOTH	11.4%	0.0%
CrossNews	BASE	8.5%	0.0%
	+FR	8.8%	0.1%
	+CO	9.1%	0.0%
	+BOTH	11.3%	0.0%

Table 11: PROMPT-CAVE MAYBE and invalid output rates across 5000 examples per variant (5 models $\times$ 1000 examples). Bold = lowest MAYBE rate per dataset.

Dataset	Variant	Invalid rate
Blog-Auth	BASE	0.5%
	+FR	0.7%
	+CO	0.6%
	+BOTH	1.2%
Fanfic	BASE	1.3%
	+FR	3.2%
	+CO	2.1%
	+BOTH	5.1%
IMDb62	BASE	0.3%
	+FR	0.4%
	+CO	0.2%
	+BOTH	0.9%
CrossNews	BASE	0.1%
	+FR	0.2%
	+CO	0.1%
	+BOTH	0.4%

Table 12: PROMPTAV invalid output rates across 5000 examples per variant (5 models $\times$ 1000 examples). Bold = lowest per dataset.

are calibration artifacts, we manually audited near-threshold gold-YES PROMPTAV outputs across all four datasets. For each model–dataset setting, we randomly sampled 30 gold-YES examples with valid PROMPTAV scores in the near-threshold score range, $0.3 \leq s \leq 0.5$. Under our fixed decision rule, these scores are mapped to NO. We then coded whether the generated rationale leaned toward different authorship, including cases where the model explicitly stated that the texts were unlikely to share an author. Table 14 shows these results. Across all settings, 92.7% of responses contained NO-leaning rationales.

Dataset	N	$\bar{\tau}^*$	$\tau = 0.5$			$\tau = \tau^*$		
			Acc	YES _A	NO _A	Acc	YES _A	NO _A
Blog-Auth	5	0.324	58.7	28.9	88.5	63.4	57.8	69.0
Fanfic	5	0.341	56.2	28.4	84.1	59.4	61.0	57.7
IMDb62	5	0.418	65.5	43.5	87.8	69.4	63.4	75.4
Cross-News	5	0.464	70.4	54.2	86.6	75.6	70.3	80.9

Table 13: Threshold sensitivity for PROMPTAV base-lines. $\bar{\tau}^*$ = mean accuracy-maximising threshold. N = number of models. Accuracy values are percentages averaged across five models.

Model	Blog	Cross	Fan	IMDb	Total
Qwen3-14B	27/30	25/30	29/30	24/30	105/120
Qwen3-32B	29/30	29/30	30/30	30/30	118/120
Gemma3-12B	29/30	30/30	30/30	30/30	119/120
Gemma3-27B	26/30	29/30	29/30	30/30	114/120
Mist-Small-24B	27/30	28/30	21/30	24/30	100/120
Overall	138/150	141/150	139/150	138/150	556/600

Table 14: Manual audit of near-threshold gold-YES PROMPTAV outputs ($0.3 \leq s \leq 0.5$). Counts indicate outputs whose natural-language rationale leaned toward NO.

C Per-Model Intervention Results

Tables 15 and 16 report the full per-model breakdown of the PROMPTAV and PROMPT-CAVE intervention results summarized in Tables 6 and 7 of the main paper.

D Additional Qualitative Examples

Table 17 provides six further false-negative correction examples complementing those in Table 9 of the main paper.

E Sensitivity Analyses: Excluding Invalid and MAYBE Outputs

Tables 18–21 report sensitivity results excluding MAYBE responses (PROMPT-CAVE) and invalid/unparseable outputs (PROMPTAV) from the denominator. The conservative NO tendency and the gains from the best-performing interventions are robust to these exclusions.

F Per-Model False-Negative Rate by Semantic Distance Quintile

Figures 2 and 3 show the false-negative rate on gold-YES pairs with valid YES/NO outputs by semantic distance quintile for each model individually, for PROMPTAV and PROMPT-CAVE interventions respectively.

Model	Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
		Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
G3-12B	BASE	0.54	0.44	0.13	0.95	0.51	0.37	0.04	0.98	0.61	0.55	0.25	0.98	0.65	0.62	0.35	0.96
	+FR	0.57	0.51	0.23	0.90	0.54	0.43	0.10	0.98	0.64	0.62	0.39	0.90	0.72	0.72	0.59	0.85
	+CO	0.55	0.47	0.17	0.94	0.51	0.38	0.05	0.98	0.59	0.53	0.23	0.96	0.63	0.61	0.39	0.87
	+BOTH	0.58	0.55	0.33	0.83	0.54	0.50	0.26	0.82	0.63	0.62	0.44	0.82	0.67	0.67	0.59	0.76
G3-27B	BASE	0.58	0.55	0.29	0.88	0.57	0.52	0.24	0.89	0.67	0.66	0.47	0.87	0.72	0.72	0.62	0.81
	+FR	0.60	0.58	0.37	0.83	0.58	0.57	0.42	0.75	0.68	0.67	0.52	0.84	0.69	0.69	0.60	0.78
	+CO	0.55	0.49	0.19	0.92	0.56	0.50	0.22	0.90	0.65	0.62	0.36	0.93	0.70	0.69	0.51	0.89
	+BOTH	0.56	0.54	0.31	0.82	0.58	0.57	0.43	0.73	0.66	0.67	0.51	0.82	0.70	0.70	0.62	0.79
Mist-24B	BASE	0.64	0.63	0.49	0.78	0.60	0.60	0.52	0.68	0.69	0.69	0.63	0.75	0.77	0.76	0.70	0.83
	+FR	0.63	0.62	0.53	0.72	0.60	0.60	0.62	0.57	0.68	0.68	0.74	0.61	0.75	0.75	0.74	0.76
	+CO	0.63	0.63	0.50	0.77	0.58	0.57	0.68	0.47	0.69	0.69	0.69	0.68	0.73	0.73	0.68	0.78
	+BOTH	0.61	0.61	0.51	0.72	0.60	0.58	0.83	0.37	0.65	0.65	0.72	0.59	0.76	0.76	0.79	0.73
Q3-14B	BASE	0.56	0.49	0.18	0.94	0.52	0.44	0.13	0.91	0.60	0.55	0.25	0.96	0.63	0.59	0.30	0.97
	+FR	0.57	0.51	0.22	0.92	0.51	0.52	0.33	0.68	0.62	0.58	0.32	0.92	0.68	0.67	0.47	0.90
	+CO	0.56	0.49	0.19	0.93	0.52	0.51	0.28	0.75	0.64	0.61	0.36	0.92	0.69	0.67	0.44	0.94
	+BOTH	0.56	0.50	0.21	0.91	0.47	0.52	0.40	0.53	0.65	0.63	0.42	0.87	0.75	0.74	0.61	0.89
Q3-32B	BASE	0.60	0.58	0.35	0.86	0.58	0.58	0.46	0.69	0.69	0.69	0.57	0.81	0.75	0.75	0.74	0.76
	+FR	0.60	0.58	0.38	0.82	0.57	0.58	0.58	0.56	0.68	0.68	0.63	0.73	0.77	0.77	0.78	0.77
	+CO	0.60	0.57	0.34	0.86	0.58	0.59	0.56	0.59	0.69	0.69	0.65	0.73	0.75	0.75	0.78	0.71
	+BOTH	0.59	0.57	0.33	0.85	0.56	0.57	0.56	0.56	0.69	0.69	0.69	0.69	0.75	0.75	0.80	0.71

Table 15: Per-model PROMPTAV intervention results. Bold = best per model–dataset column. Averaged results in Table 6.

Model	Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
		Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
G3-12B	BASE	0.53	0.42	0.08	0.98	0.50	0.34	0.00	1.00	0.54	0.42	0.09	0.99	0.61	0.56	0.25	0.98
	+FR	0.58	0.52	0.22	0.95	0.52	0.39	0.05	0.98	0.64	0.61	0.34	0.95	0.71	0.70	0.52	0.90
	+CO	0.53	0.41	0.08	0.98	0.50	0.34	0.00	1.00	0.52	0.38	0.04	0.99	0.61	0.54	0.23	0.99
	+BOTH	0.57	0.50	0.19	0.95	0.51	0.37	0.03	0.99	0.60	0.54	0.23	0.98	0.68	0.66	0.43	0.93
G3-27B	BASE	0.62	0.59	0.36	0.88	0.54	0.44	0.12	0.96	0.72	0.70	0.47	0.96	0.78	0.78	0.66	0.91
	+FR	0.64	0.63	0.46	0.82	0.58	0.57	0.37	0.80	0.78	0.78	0.71	0.86	0.80	0.80	0.80	0.79
	+CO	0.60	0.56	0.30	0.89	0.53	0.40	0.07	0.98	0.69	0.66	0.39	0.98	0.78	0.78	0.65	0.92
	+BOTH	0.63	0.62	0.45	0.81	0.60	0.59	0.37	0.83	0.78	0.79	0.69	0.88	0.80	0.80	0.80	0.80
Mist-24B	BASE	0.59	0.60	0.39	0.79	0.57	0.60	0.61	0.54	0.64	0.67	0.63	0.64	0.77	0.79	0.82	0.73
	+FR	0.59	0.60	0.41	0.76	0.54	0.57	0.63	0.46	0.58	0.64	0.61	0.55	0.74	0.76	0.86	0.62
	+CO	0.53	0.50	0.19	0.87	0.53	0.54	0.37	0.69	0.60	0.62	0.40	0.81	0.74	0.77	0.71	0.78
	+BOTH	0.56	0.55	0.28	0.84	0.55	0.57	0.62	0.47	0.60	0.63	0.42	0.78	0.74	0.77	0.80	0.69
Q3-14B	BASE	0.47	0.41	0.07	0.86	0.39	0.36	0.05	0.72	0.46	0.42	0.07	0.86	0.51	0.50	0.12	0.90
	+FR	0.49	0.45	0.13	0.85	0.42	0.43	0.15	0.68	0.50	0.52	0.19	0.80	0.55	0.62	0.28	0.82
	+CO	0.49	0.41	0.08	0.90	0.42	0.35	0.03	0.81	0.48	0.43	0.09	0.88	0.51	0.50	0.12	0.90
	+BOTH	0.49	0.42	0.09	0.89	0.37	0.40	0.11	0.64	0.47	0.48	0.13	0.81	0.47	0.54	0.18	0.77
Q3-32B	BASE	0.59	0.59	0.37	0.82	0.59	0.61	0.57	0.61	0.71	0.74	0.69	0.74	0.81	0.84	0.79	0.83
	+FR	0.59	0.61	0.46	0.71	0.57	0.59	0.65	0.49	0.70	0.72	0.74	0.66	0.82	0.84	0.85	0.78
	+CO	0.56	0.56	0.30	0.82	0.56	0.59	0.49	0.64	0.69	0.71	0.58	0.79	0.79	0.82	0.73	0.84
	+BOTH	0.57	0.58	0.37	0.78	0.54	0.58	0.55	0.53	0.68	0.71	0.64	0.72	0.81	0.84	0.80	0.81

Table 16: Per-model PROMPT-CAVE intervention results. Bold = best per model–dataset column. Averaged results in Table 7.

Setting	Outcome	Baseline rationale excerpt	Intervention rationale excerpt
BLOG-AUTH G3-27B PROMPTAV +FR row 361	Gold: YES Base: NO Interv.: YES	“However, the substantial differences in almost all other areas (tone, sentence structure, punctuation, capitalization, and content) strongly suggest different authors.”	“The shared characteristics point to a writer who is comfortable with very informal language, fragmented sentences, and a direct, conversational tone. While the content differs, the way it’s written is remarkably similar.”
FANFIC Q3-14B PROMPTAV +BOTH row 494	Gold: YES Base: NO Interv.: YES	“These differences suggest that the authorship is unlikely to be the same, especially given the distinct genres (fantasy vs. shonen) and different character dynamics.”	“While the content and specific expressions differ, the surface-form habits are consistent enough to suggest a high likelihood that the texts were written by the same author.”
IMDB62 Mist-24B PROMPT-AVE +FR row 924	Gold: YES Base: NO Interv.: YES	“Text 1 focuses on a video game, while Text 2 critiques a film. The subject matter and purpose are entirely different.”	“Both texts use exclamatory phrases and rhetorical questions. The use of humor and sarcasm is also present in both.”
CROSSNEWS Mist-24B PROMPTAV +CO row 427	Gold: YES Base: NO Interv.: YES	“Both texts are about tourism but cover different subtopics. The vocabulary and subject matter are distinct enough to suggest different authors or at least different reporting angles.”	“The stylometric evidence strongly suggests the same author, but the different content slightly reduces the confidence.”
FANFIC G3-27B PROMPTAV +BOTH row 890	Gold: YES Base: NO Interv.: YES	“The similarities are enough to suggest some possibility of a shared author, but the significant differences in content and stylistic focus make it unlikely.”	“The low-level stylometric evidence suggests a moderate likelihood of shared authorship.”
CROSSNEWS Q3-14B PROMPTAV +CO row 322	Gold: YES Base: NO Interv.: YES	“Subject matter is completely different, which is a strong indicator that the texts were written by different authors.”	“The low-level stylometric features are very consistent, suggesting a high likelihood of the same author.”

Table 17: Additional false-negative corrections across four datasets where the intervention shifts focus to surface stylistic habits and recovers the gold YES label.

Model	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
G3-12B	0.54	0.42	0.08	0.98	0.50	0.34	0.00	1.00	0.54	0.43	0.09	0.99	0.63	0.57	0.26	0.98
G3-27B	0.62	0.59	0.36	0.88	0.54	0.44	0.12	0.97	0.72	0.71	0.48	0.97	0.79	0.79	0.67	0.91
Mist-24B	0.65	0.63	0.44	0.84	0.62	0.62	0.65	0.59	0.70	0.70	0.69	0.72	0.81	0.81	0.88	0.75
Q3-14B	0.58	0.45	0.10	0.99	0.56	0.42	0.08	0.98	0.62	0.48	0.12	1.00	0.72	0.60	0.24	0.99
Q3-32B	0.64	0.62	0.41	0.86	0.64	0.64	0.62	0.65	0.77	0.77	0.75	0.79	0.86	0.86	0.86	0.86

Table 18: PROMPT-CAVE baseline performance excluding MAYBE outputs. Acc = accuracy, F1 = macro-F1, YES_A/NO_A = per-class accuracy. Bold = best per dataset column.

Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
BASE	0.60	0.54	0.28	0.91	0.57	0.49	0.30	0.84	0.67	0.62	0.43	0.89	0.76	0.72	0.58	0.90
+FR	0.63	0.59	0.37	0.87	0.58	0.54	0.41	0.75	0.71	0.69	0.58	0.83	0.79	0.79	0.74	0.83
+CO	0.58	0.50	0.21	0.93	0.55	0.46	0.22	0.88	0.65	0.59	0.34	0.94	0.76	0.72	0.55	0.92
+BOTH	0.61	0.56	0.30	0.90	0.59	0.53	0.38	0.78	0.71	0.67	0.48	0.90	0.79	0.77	0.69	0.86

Table 19: PROMPT-CAVE intervention results excluding MAYBE outputs, averaged across five models. Bold = best per dataset column.

Model	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
G3-12B	0.54	0.44	0.13	0.95	0.51	0.37	0.04	0.98	0.61	0.55	0.25	0.98	0.65	0.62	0.35	0.96
G3-27B	0.59	0.55	0.29	0.89	0.57	0.52	0.24	0.89	0.68	0.66	0.47	0.89	0.72	0.72	0.63	0.81
Mist-24B	0.64	0.63	0.49	0.78	0.60	0.60	0.52	0.68	0.69	0.69	0.63	0.75	0.77	0.76	0.70	0.83
Q3-14B	0.56	0.49	0.18	0.94	0.54	0.45	0.13	0.94	0.60	0.55	0.25	0.96	0.64	0.59	0.30	0.97
Q3-32B	0.61	0.58	0.35	0.86	0.59	0.59	0.47	0.71	0.69	0.69	0.58	0.81	0.75	0.75	0.74	0.76

Table 20: PROMPTAV baseline performance excluding invalid outputs. Acc = accuracy, F1 = macro-F1, YES_A/NO_A = per-class accuracy. Bold = best per dataset column.

Variant	Blog-Auth				Fanfic				IMDb62				Cross-News			
	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A	Acc	F1	YES _A	NO _A
BASE	0.59	0.54	0.29	0.88	0.56	0.51	0.28	0.84	0.66	0.63	0.44	0.88	0.70	0.69	0.54	0.87
+FR	0.60	0.57	0.35	0.85	0.58	0.55	0.43	0.73	0.66	0.65	0.52	0.80	0.72	0.72	0.64	0.81
+CO	0.58	0.53	0.28	0.89	0.56	0.52	0.37	0.75	0.65	0.63	0.46	0.85	0.70	0.69	0.56	0.84
+BOTH	0.59	0.56	0.34	0.84	0.58	0.56	0.53	0.63	0.66	0.65	0.56	0.77	0.73	0.73	0.68	0.78

Table 21: PROMPTAV intervention results excluding invalid outputs, averaged across five models. Bold = best per dataset column.

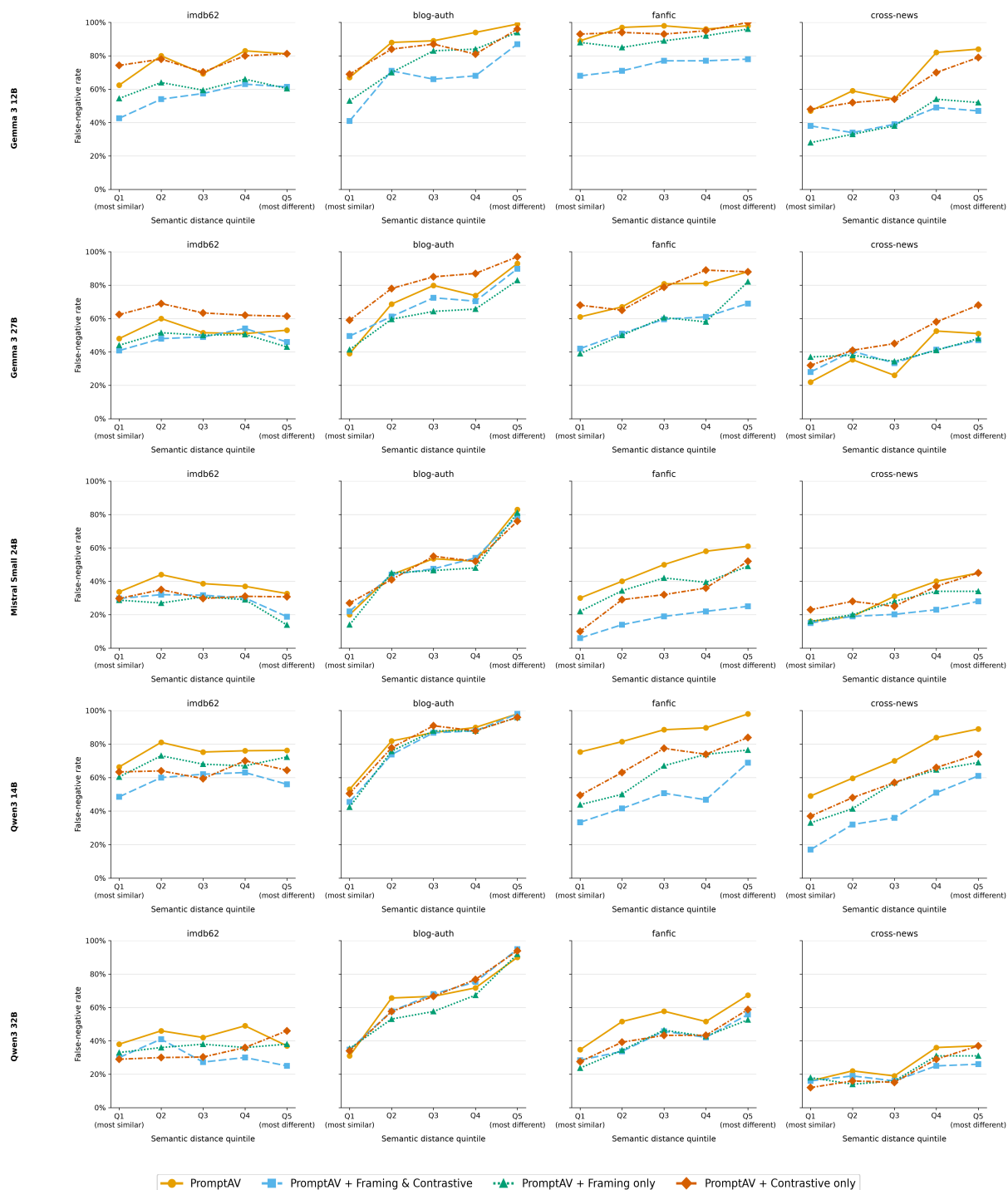


Figure 2: Per-model false-negative rate on gold-YES pairs with valid YES/NO outputs by semantic distance quintile for PROMPTAV interventions.

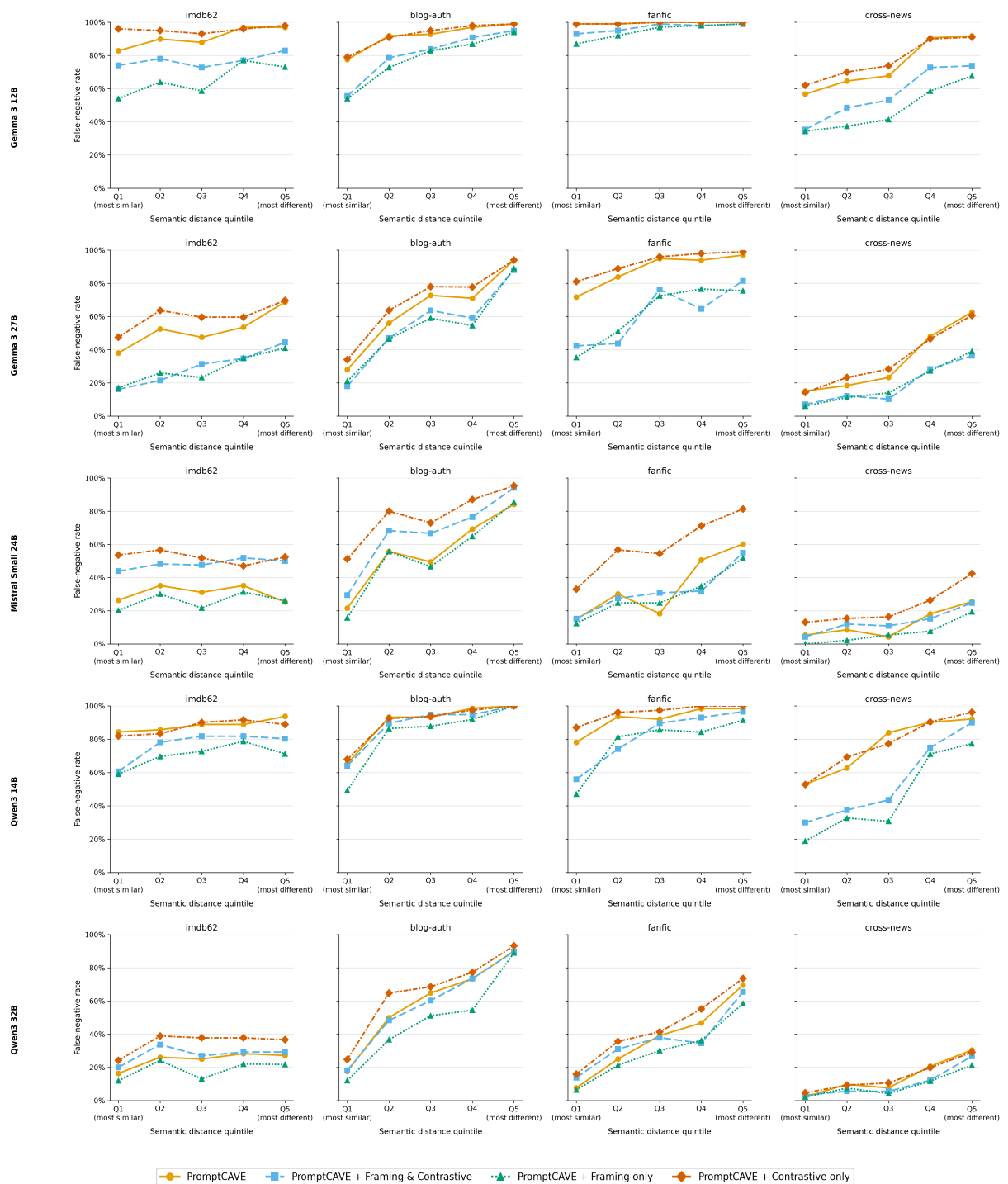


Figure 3: Per-model false-negative rate on gold-YES pairs with valid YES/NO outputs by semantic distance quintile for PROMPT-CAVE interventions.

G Prompt Templates

Table 22: Zero-shot prompt templates used in the experiments. The placeholders {text1} and {text2} are replaced with the paired input texts.

Component / Method	Prompt text or template composition
Shared prompt components	
Score task	Task: On a scale of 0 to 1, with 0 indicating low confidence and 1 indicating high confidence, please provide a general assessment of the likelihood that Text 1 and Text 2 were written by the same author. Your answer should reflect a moderate level of strictness in scoring.
COT suffix	Let's think step by step.
Framing instruction	Do not focus on what the texts are about. Focus on recurring surface-form habits that could remain stable across topics.
Contrastive instruction	When comparing the texts, explicitly separate low-level stylometric evidence from content-sensitive evidence, and give more weight to low-level stylometric evidence.
Stylometric variables	Here are some relevant variables to this problem. 1. punctuation style(e.g. hyphen, brackets, colon, comma, parenthesis, quotation mark) 2. special characters style, capitalization style(e.g. Continuous capitalization, capitalizing certain words) 3. acronyms and abbreviations(e.g. Usage of acronyms such as OMG, Abbreviations without punctuation marks such as Mr Rochester vs. Mr. Rochester, Unusual abbreviations such as def vs. definitely) 4. writing style 5. expressions and Idioms 6. tone and mood 7. sentence structure 8. any other relevant aspect
PromptAV reasoning instruction	First step: Understand the problem, extracting relevant variables and devise a plan to solve the problem. Then, carry out the plan and solve the problem step by step. Finally, show the confidence score.
Prompt-CAVE output format	Provide the answer in a Python JSON format. Use the following keys for your dictionary: 'punctuation style', 'special characters style, capitalization style', 'acronyms and abbreviations', 'writing style', 'expressions and Idioms', 'tone and mood', 'sentence structure', 'any other relevant aspect', 'final score'. Apart from the 'final score', everything else must have a text value; also, the text should include a concluding YES/NO/MAYBE about whether the two texts are similar or not with respect to the key at hand. Finally, provide an 'output' key in your dictionary, which says YES if the two texts are written by the same author, and NO otherwise.
Input block	Text 1: {text1} Text 2: {text2}
Method-specific prompt templates	
COT	Score task + COT suffix + Input block.
PROMPTAV	Score task + Stylometric variables + PromptAV reasoning instruction + Input block.
PROMPTAV + Framing	Score task + Framing instruction + Stylometric variables + PromptAV reasoning instruction + Input block.
PROMPTAV + Contrastive	Score task + Stylometric variables + Contrastive instruction + PromptAV reasoning instruction + Input block.
PROMPTAV + Framing + Contrastive	Score task + Framing instruction + Stylometric variables + Contrastive instruction + PromptAV reasoning instruction + Input block.
PROMPT-CAVE	Score task + Stylometric variables + Prompt-CAVE output format + Input block.
PROMPT-CAVE + Framing	Score task + Framing instruction + Stylometric variables + Prompt-CAVE output format + Input block.
PROMPT-CAVE + Contrastive	Score task + Stylometric variables + Contrastive instruction + Prompt-CAVE output format + Input block.
PROMPT-CAVE + Framing + Contrastive	Score task + Framing instruction + Stylometric variables + Contrastive instruction + Prompt-CAVE output format + Input block.