Beyond Representational Similarity: Source-Conditioned Description-Length Gain for Generative Plagiarism Detection and Candidate Source Reranking

Peijia Guo[1]*, Wenxuan Xie[1]*, ZiGuang Li[1]*, Ming Li[2]†

[1] Shanghai Institute for Mathematics and Interdisciplinary Sciences, Fudan University
[2] University of Waterloo
24114020005@m.fudan.edu.cn, mli@uwaterloo.ca

Abstract

Large language models (LLMs) pose challenges to academic integrity and peer review. Yet generative plagiarism detection remains an underexplored and largely unresolved challenge. Prior work on LLM-generated-text detection targets AI involvement, which may be permissible, rather than source reuse, while similarity-based methods struggle after extensive rewriting and multi-source synthesis. Motivated by the description-length view of probabilistic prediction, in which relevant side information can reduce a target sequence's code length, we introduce Source-Conditioned Description-Length Gain (SCDG), a directional, training-free framework that contrasts a frozen language model's description length of a suspicious document P with and without a candidate source S. This contrast yields token-level log-likelihood gains that measure the incremental predictive evidence supplied by S. We evaluate SCDG on the PAN at CLEF benchmarks for generative plagiarism. On a PAN 2025-derived pairwise benchmark, SCDG achieves 0.92 Precision, 0.97 Recall, and 0.94 F1, outperforming all baselines; on PAN 2026's multi-source retrieval task, it reaches 0.83 nDCG@10 and 0.96 Recall@100, surpassing all baselines. On a same-topic, same-event Multi-News test, the calibrated gain-distribution SCDG classifier predicts source reuse for only 0.125% of pairs, supporting robustness to topical overlap under this evaluation protocol. These results establish SCDG as a unified and token-decomposable signal for source-specific content reuse under extensive transformation.

Introduction

As LLMs evolve from writing assistants into research agents, they are becoming embedded in scholarly knowledge production. They can retrieve sources, synthesize evidence, organize arguments, and draft reports with limited human intervention. These capabilities enable deeper forms of rewriting, in which the central ideas, evidence, reasoning patterns, and organizational logic of source materials are recombined into fluent new prose with little recognizable lexical overlap. A document may therefore depend heavily on a particular source while bearing limited surface resemblance to it. Detecting this form of *generative plagiarism* requires identifying source reuse under extensive transformation.

LLM-generated-text detection does not address the same question. Existing classifiers and likelihood-based detectors estimate whether an LLM contributed to a text (Yang et al. 2024; Mitchell et al. 2023; Bao et al. 2024), whereas generative plagiarism detection asks whether the text reuses content from a particular source. AI involvement neither implies nor is required for such reuse: machine-generated text may be independently produced, while source-derived content may remain after substantial human revision.

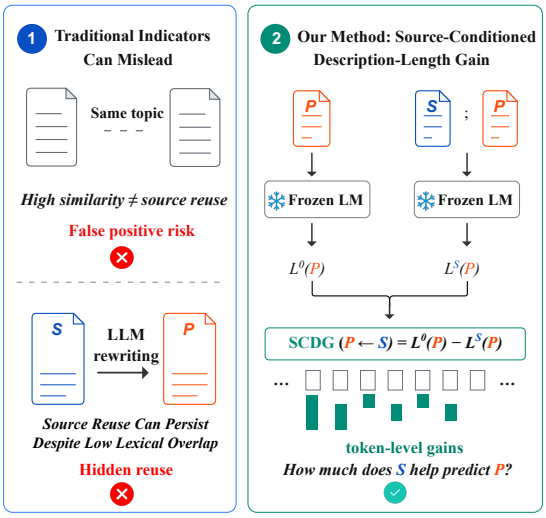

Figure 1: SCDG distinguishes source-conditioned predictive evidence from surface similarity under topical overlap and extensive rewriting.

Classical external plagiarism detection typically combines source retrieval with passage alignment (Potthast et al. 2010; Foltýnek, Meuschke, and Gipp 2019). Lexical matching and BM25 work well when reused passages retain recognizable wording, while embedding-based methods extend the comparison to semantic similarity. Both signals, however, weaken under extensive paraphrasing, reorganization, and multi-source synthesis (Wahle et al. 2021, 2022; Greiner-Petter et al. 2025). Prompted LLMs can also judge document pairs directly (Lee et al. 2025), but their decisions may depend on prompting and do not yield a stable measure of the contribution made by each candidate source.

The remaining difficulty is that resemblance does not reliably indicate source dependence. Independently written documents about the same topic may be highly similar, whereas

*These authors contributed equally.
†Corresponding author.

a deeply rewritten document may be only weakly similar to the source on which it relies. We therefore seek a directional, source-specific signal that measures how much a candidate source S contributes to a suspicious document P, while controlling for the baseline predictability of P. This motivates a different view of source evidence based on description length. Under a fixed probabilistic model, the negative log-probability of a sequence corresponds to its model-relative codelength (Shannon 1948; Rissanen 1978; MacKay 2003). We compare the codelength assigned to the same suspicious document with and without a candidate source in the model context. When the document reuses information from that source, conditioning on it should make the reused content easier to predict and thereby shorten the document's codelength. The resulting reduction measures the incremental predictive contribution of the candidate source.

Building on this idea, we introduce *Source-Conditioned Description-Length Gain* (SCDG), a training-free framework that measures the predictive contribution of a candidate source, as illustrated in Figure 2. SCDG compares the description length assigned by a frozen language model to the same suspicious document with and without the source in context. A larger reduction indicates that the source provides information that helps the model predict the observed text. Because this reduction decomposes over individual target tokens, SCDG also reveals where source conditioning provides predictive evidence and allows sparse source-derived content to be aggregated without being diluted by the rest of the document. Unlike lexical or embedding similarity, SCDG measures an asymmetric relation from a candidate source to a suspicious document while controlling for the document's baseline predictability. The resulting score can be used both to classify suspicious–source pairs and to rerank candidate sources retrieved from a large corpus, without task-specific parameter updates. SCDG should be interpreted as model-relative evidence of source dependence rather than direct proof of provenance or authorial intent.

We evaluate SCDG on both pairwise detection and full-corpus source retrieval using the PAN generative-plagiarism benchmarks. On the PAN 2025-derived pairwise benchmark, SCDG performs consistently across three frozen language models and reaches an F_1 score of 0.9433, outperforming lexical, embedding-based, and prompted-LLM baselines. On PAN 2026, SCDG substantially improves two matched candidate rankings, raising nDCG@10 by more than 0.13 without changing their candidate sets. These results show that source-conditioned predictability provides a robust signal for detecting transformed source reuse and complements conventional retrieval methods in large-scale source attribution. We further conduct two targeted analyses of the SCDG signal. A controlled source-retention study shows that SCDG increases as more genuine evidence is retained for reused content, while remaining comparatively stable on newly written content. On a complementary Multi-News test set, SCDG-based classifier produces positive decisions for only 0.125% of same-topic, same-event article pairs, indicating limited susceptibility to topical confounding under this evaluation protocol. Together, these analyses examine both SCDG's sensitivity to genuine source evidence and its ability to distinguish source dependence from topical overlap.

Our main contributions are as follows:

- We approach generative plagiarism detection through the lens of source-conditioned predictability and introduce *Source-Conditioned Description-Length Gain* (SCDG), a training-free, token-level framework that measures the predictive contribution of a candidate source to a suspicious document.
- We develop DAAC, an uncertainty-aware retrieval-fusion method that combines lexical, sentence-level dense, and abstract-level evidence using retrieval ranks alone. DAAC constructs compact candidate sets without additional neural inference and, together with SCDG, enables scalable retrieval and attribution of potential source documents from large corpora.
- We provide a comprehensive empirical evaluation of SCDG. Our method outperforms all evaluated baselines on both the PAN 2025-derived and PAN 2026 benchmarks, while source-retention and Multi-News experiments further support its sensitivity to genuine source evidence and robustness to topical confounding.

Related Work

External plagiarism detection. Classical external plagiarism detection treats the task as source retrieval followed by passage alignment. PAN established standardized corpora and evaluation measures for this setting (Potthast et al. 2010); subsequent editions separated retrieval from detailed comparison and evaluated increasingly realistic paraphrase, translation, and summarization cases (Potthast et al. 2011, 2012, 2013, 2014). These benchmarks showed strong performance on verbatim reuse but persistent difficulty with heavily obfuscated reuse. Most such systems nevertheless decide reuse through explicit surface or representational similarity. This signal weakens when rewriting removes local correspondence or when evidence is distributed across multiple sources. Our method instead asks whether a candidate source makes the observed suspicious text more predictable under a fixed autoregressive model.

Generative plagiarism detection. Neural paraphrasers can maintain high semantic similarity to source texts while making paraphrase detection substantially more difficult (Wahle et al. 2021); with larger autoregressive language models, GPT-3 paraphrases were nearly indistinguishable from original texts to human judges (53% accuracy), while the best tested model achieved only about 66% macro-F1 in detecting them (Wahle et al. 2022). A study extending language-model memorization analysis further identifies verbatim, paraphrase, and idea-level reuse from training data, but its high-precision retrieval-and-alignment pipeline yields conservative lower-bound estimates because of limited recall (Lee et al. 2023). PlagBench expands benchmark coverage with 46.5K synthetic text pairs spanning verbatim copying, paraphrasing, and summarization, and demonstrates strong plagiarism-detection performance for GPT-4 Turbo, although its advantage varies across tasks and baselines (Lee et al.

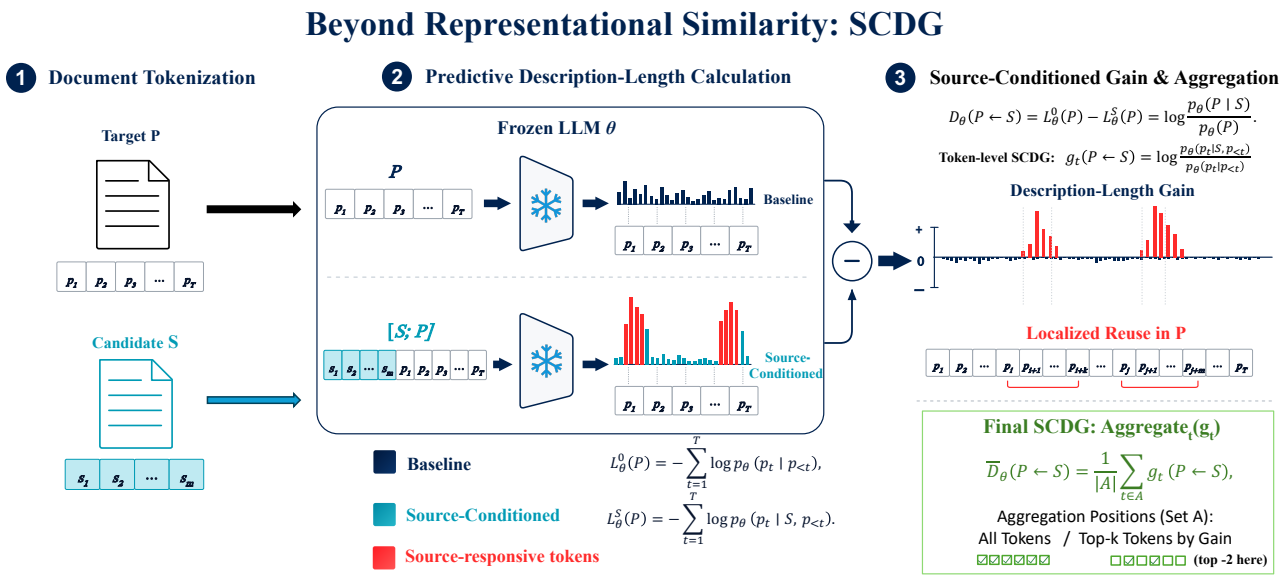

Figure 2: Overview of Source-Conditioned Description-Length Gain (SCDG). Given a suspicious document P and a candidate source S, both documents are tokenized, and a frozen autoregressive language model scores the identical target-token sequence P under two contexts: the target prefix alone and the source-conditioned context $[S; P]$. The resulting baseline and source-conditioned codelengths define the total description-length gain $D_\theta(P \leftarrow S) = L_\theta^0(P) - L_\theta^S(P)$, which decomposes exactly into token-level gains $g_t(P \leftarrow S)$. Positive gains identify source-responsive positions at which conditioning on S makes the observed target tokens more predictable, thereby revealing localized source reuse. Finally, SCDG averages the gains over all target tokens or the top-k tokens to obtain the length-normalized score $\overline{D}_\theta(P \leftarrow S)$ for pairwise plagiarism detection and candidate-source reranking.

2025). However, such prompted LLM judgments are sensitive to prompting and do not directly quantify the incremental contribution of a candidate source to the observed target. PAN 2025 reported promising paragraph-level performance for embedding-based systems, reaching about 0.8 recall at roughly 0.5 precision, but also revealed weak generalization to earlier PAN data (Greiner-Petter et al. 2025). PAN 2026 moves to a more realistic one- or multi-source generation setting, placing greater emphasis on robustness to deep rewriting, multi-source synthesis, and distribution shift (Bevendorff et al. 2026).

LLM-generated-text detection. A separate literature detects AI authorship using supervised classifiers, zero-shot likelihood statistics, or watermarking (Yang et al. 2024). DetectGPT obtains strong zero-shot discrimination from probability curvature (Mitchell et al. 2023), Fast-DetectGPT substantially improves its efficiency and accuracy through conditional curvature (Bao et al. 2024), and ImBD targets machine-revised text by aligning to machine stylistic preferences (Chen et al. 2025). These methods detect LLM involvement rather than source-specific reuse, even though LLM assistance does not necessarily imply plagiarism. SCDG instead measures the model-relative predictive evidence supplied by a candidate source.

Methodology

Theoretical Foundation: Conditional Description Length

According to Shannon's source coding theory and its MDL interpretation, the negative log-probability of a sequence under a fixed probabilistic model corresponds to its ideal model-relative codelength; a lossless code can realize this length up to an additive coding constant (Shannon 1948; Rissanen 1978; MacKay 2003). Near-optimal compression of natural language requires more than modeling surface statistics or merely representing semantic content; it requires capturing the deeper generative regularities underlying linguistic structure and meaning (Shannon 1951; Mahoney 1999; Delétang et al. 2024; Li et al. 2025). From an algorithmic-information-theoretic perspective, the conditional codelength induced by a fixed computable model provides a computable, model-relative upper bound on conditional Kolmogorov complexity, up to an additive model-dependent constant (Li and Vitányi 2008). Consequently, a substantial reduction in conditional codelength indicates that, conditioning on the candidate source yields a model-relative incremental information gain for predicting the target. This signal allows our framework to remain informative despite surface-form changes introduced by LLM paraphrasing and capture source-derived information that survives rewriting. Using natural logarithms, we

define the unconditional codelength $L_\theta^0(P)$ and the source-conditioned codelength $L_\theta^S(P)$ of document P, measured in nats, as:

$$L_\theta^0(P) = -\sum_{t=1}^{T} \log p_\theta(p_t \mid p_{<t}), \tag{1}$$

$$L_\theta^S(P) = -\sum_{t=1}^{T} \log p_\theta(p_t \mid S, p_{<t}). \tag{2}$$

The total description-length gain yielded by the candidate source S is quantified by their reduction:

$$D_\theta(P \leftarrow S) = L_\theta^0(P) - L_\theta^S(P) = \log \frac{p_\theta(P \mid S)}{p_\theta(P)}. \tag{3}$$

Mathematically, $D_\theta(P \leftarrow S)$ is a model-relative conditional log-likelihood ratio and can be viewed as a PMI-like information-density quantity under p_θ. Our contribution lies not in this algebraic identity alone, but in its operationalization as a directional source-reuse signal, its exact tokenwise decomposition and sparse evidence aggregation, its unified use for pairwise detection and candidate-source reranking, and its empirical validation under extensive rewriting and multi-source retrieval.

In autoregressive language modeling, negative log-likelihood quantifies how surprising, and hence how difficult to predict, an observed sequence is under the model. Therefore, D_θ intuitively quantifies how much easier it is for the model to predict the observed document P when guided by source S. When S is provided in the prompt context, it can act as a probabilistic generative blueprint. If P reuses the ideas, organization, or reasoning patterns of S, the model can exploit this source-provided generative logic rather than predict P solely from its own prefix, potentially causing the codelength of P to drop substantially.

Validating the Theoretical Intuition through Controlled Source-Evidence Retention

The preceding analysis connects conditional codelength reduction to source-guided predictability, but does not establish whether the observed gain is driven by genuinely reused source evidence. We therefore conduct a controlled source-evidence intervention on 300 aligned suspicious–source pairs from the PAN 2025 validation corpus. For each pair, we construct a source context $S^{(r)}$ retaining a proportion $r \in 0, 0.25, 0.50, 0.75, 1$ of the annotated source evidence, replacing removed chunks with length-matched distractor passages randomly sampled from an external news corpus. Each retention level is evaluated over five random seeds under otherwise fixed scoring conditions, with length-normalized gains averaged separately for plagiarism-labeled and new content.

As shown in Figure 3, plagiarism-labeled content exhibits a strong and monotonic response to retained source evidence across Qwen3-8B-Base(Yang et al. 2025), Llama-3.1-8B(Grattafiori et al. 2024), and Ministral-3-8B-Base-2512(Liu et al. 2026). In contrast, the gain of new content remains close to zero across all retention levels and shows little response to increasing retention of genuine source evidence. This consistent separation supports our design premise: description-length gain is selectively sensitive to genuinely reused source evidence and can therefore serve as an effective signal for generative plagiarism detection.

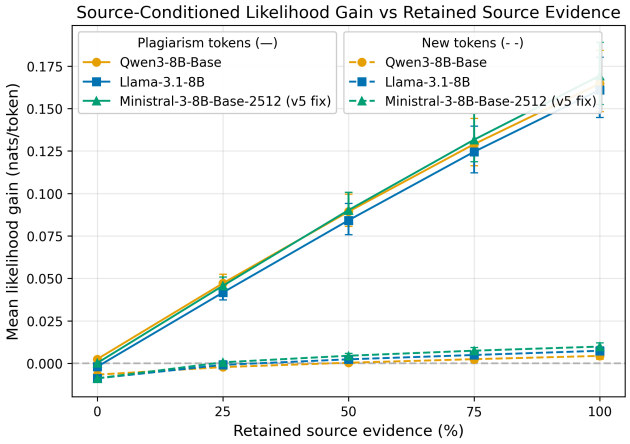

Figure 3: Mean length-normalized description-length gain over plagiarism-labeled and new content as a function of the retained source-evidence ratio. Error bars denote 95% confidence intervals. Across all three frozen language models, plagiarism-labeled content exhibits a strong monotonic response to genuine source evidence, whereas gains for new content remain close to zero.

Source-Conditioned Description-Length Gain (SCDG)

The pairwise codelength contrast is model-relative and does not assume that p_θ induces a universal or optimal code. Both likelihood terms are evaluated by the same frozen model over the same observed target-token sequence, differing only in whether the candidate source S is absent or available as side information.

Because autoregressive codelength is additive over target positions, the document-level gain decomposes exactly into tokenwise contributions. We define the token-level SCDG as

$$g_t(P \leftarrow S) = \log \frac{p_\theta(p_t \mid S, p_{<t})}{p_\theta(p_t \mid p_{<t})}, \tag{4}$$

$$D_\theta(P \leftarrow S) = \sum_{t=1}^{T} g_t(P \leftarrow S). \tag{5}$$

For a selected set of target positions $A \subseteq \{1, \ldots, T\}$, the corresponding length-normalized gain is measured in nats per token.

$$\overline{D}_\theta(P \leftarrow S) = \frac{1}{|A|} \sum_{t \in A} g_t(P \leftarrow S). \tag{6}$$

This formulation supports exploratory inspection of token-level contributions as well as document-level aggregation; the selection of A used for detection is specified in the experimental setup. Specifically, A contains either all target-token

positions for document-wide averaging or the positions of the top-k gains for sparse-evidence aggregation.

A positive $g_t(P \leftarrow S)$ means that conditioning on S increases the probability assigned by the model to the observed token p_t, thereby shortening its codelength. The resulting sequence of tokenwise gains provides a directional evidence profile indicating where the candidate source improves prediction of the suspicious document.

Pairwise Detection and Candidate-Source Reranking

We use $\overline{D}_\theta(P \leftarrow S)$ to denote the resulting scalar score. This unified score supports both pairwise detection and candidate-source reranking without task-specific fine-tuning or parameter updates.

For pairwise plagiarism detection, we predict whether P reuses content from S by thresholding their directional description-length gain:

$$\widehat{y}(P, S) = \mathbb{I}\big[\overline{D}_\theta(P \leftarrow S) > \tau\big]. \tag{7}$$

where τ is calibrated on validation data.

For multi-source retrieval and attribution, a first-stage lexical or dense retriever returns a candidate set $\mathcal{C}(P)$ of size K. We then rank the candidate sources in descending order of their description-length gains:

$$\pi_P = \operatorname{argsort}^{\downarrow}_{S \in \mathcal{C}(P)} \overline{D}_\theta(P \leftarrow S). \tag{8}$$

where π_P denotes the resulting ranked list. Candidates that yield larger source-conditioned reductions in the codelength of P receive higher ranks. In this way, the same directional and token-decomposable signal supports both binary source-reuse detection and training-free candidate-source reranking under extensive generative rewriting.

Dense-Anchored Abstract Calibration (DAAC) indicator. For a fixed query q, let $r_j(x)$ denote the rank of candidate x under route $j \in \{B, D, A\}$, corresponding to BM25, sentence-level dense, and abstract-level retrieval. We convert each rank into reciprocal-rank evidence:

$$e_j(x) = \frac{\mathbb{I}[r_j(x) < \infty]}{k + r_j(x)}, \qquad j \in \{B, D, A\}, \tag{9}$$

where $k = 60$ and missing candidates receive zero evidence. For readability, we write $b(x) = e_B(x)$, $d(x) = e_D(x)$, and $a(x) = e_A(x)$. The dense-route uncertainty and abstract-route confidence are

$$\begin{aligned} U_D(x) &= 1 - \min\{1, (k+1)d(x)\}, \\ C_A(x) &= \frac{\mathbb{I}[r_A(x) < \infty]}{\log_2(r_A(x) + 1)}. \end{aligned} \tag{10}$$

The DAAC relevance indicator is

$$\begin{aligned} s_{\mathrm{DAAC}}(q, x) = d(x)\big(1 + U_D(x) C_A(x)\big) \\ + b(x)^2 + a(x)^2. \end{aligned} \tag{11}$$

The gated interaction allows abstract confidence to reinforce uncertain dense evidence, while the squared BM25 and abstract terms provide weak complementary corrections without dominating the dense route.

Because DAAC depends only on retrieval ranks, it requires no additional neural inference. From the union of the three retrieval lists, $\mathcal{U}_q$, we retain

$$\mathcal{C}_q = \mathrm{TopK}_{x \in \mathcal{U}_q}\, s_{\mathrm{DAAC}}(q, x), \qquad K = 1000, \tag{12}$$

and apply the more expensive SCDG scoring only to $\mathcal{C}_q$. After query-wise min–max normalization $\mathcal{N}_q(\cdot)$, the final score is

$$\begin{aligned} s_{\mathrm{final}}(q, x) = (1 - \lambda)\, \mathcal{N}_q\big(s_{\mathrm{DAAC}}(q, x)\big) \\ + \lambda\, \mathcal{N}_q\big(s_{\mathrm{ours}}(q, x)\big), \qquad x \in \mathcal{C}_q, \end{aligned} \tag{13}$$

where $\lambda \in [0, 1]$ is selected on the development folds. Candidates are ranked in descending order of s_{final}.

Experimental Setup

We evaluate SCDG through three questions: **Q1** whether it provides a consistent pair-level signal across frozen language-model backends; **Q2** whether sparse aggregation improves full-corpus candidate-source ranking; and **Q3** whether the signal is confounded by same-topic relatedness.

Datasets. We evaluate SCDG in three settings: pairwise source-reuse detection, full-corpus source retrieval, and same-topic confounding. The PAN25-derived benchmark contains 6,791 scientific source–suspicious document pairs, including 4,777 positives and 2,014 negatives. We split the data by document-linked groups to prevent document overlap across partitions. PAN26 evaluates open-corpus multi-source retrieval with 200 suspicious queries, 86,822 candidate sources, and 614 graded query–source relevance judgments. To evaluate susceptibility to topical confounding, we construct 8,000 same-cluster article pairs from Multi-News and test whether a method assigns source-reuse decisions merely because two documents cover the same topic or event. We use a cluster-disjoint 6,400/1,600 train–test split.

Baselines. We compare SCDG with lexical, embedding, prompted-LLM, and retrieval-fusion baselines.

Pairwise generative plagiarism detection. For Q1, we evaluate (1) **PAN12-style character-n-gram matching** for normalized lexical overlap (Potthast et al. 2012); (2) **BM25 pair scoring** for term-weighted lexical relevance (Robertson and Zaragoza 2009); (3) **Linq-Embed-Mistral cosine similarity**, adapted to document-pair scoring from the strongest post-hoc embedding baseline in the official PAN 2025 overview (Choi et al. 2024; Greiner-Petter et al. 2025); and (4) **PlagBench-style LLM judging** with Meta-Llama-3-8B-Instruct under zero- and few-shot prompting, with and without chain-of-thought (Lee et al. 2025).

Full-corpus candidate-source retrieval. For Q2, the six baselines comprise **BM25-full** and **BM25-sentence**, which use full-document and sentence-window query units over the same document-level index; **BGE-M3 Dense-full** and **BGE-M3 Dense-sentence** at the corresponding query granularities (Chen et al. 2024); **Four-way CombSUM**, which combines normalized scores from these four routes (Fox and Shaw 1994); and **Linq-Embed-Mistral sentence-to-chunk retrieval**. We further evaluate SCDG reranking on the top 1,000

Method		*Threshold-based* Prec.	Rec.	F_1	FPR↓	BAcc.	MCC	*Threshold-free* AUROC	AP
Baselines									
PAN12-style n-gram overlap		0.8545	0.9602	0.9041	0.3861	0.7874	0.6416	0.8482	0.8842
BM25 pair score		0.7483	0.9539	0.8369	0.7587	0.5977	0.2826	0.7093	0.8237
Linq-Embed-Mistral cosine		0.7729	0.9497	0.8515	0.6597	0.6449	0.3874	0.7904	0.8813
PlagBench zero-shot (vanilla)		0.7032	**1.0000**	0.8258	0.9975	0.5012	0.0417	0.5476	0.7250
PlagBench zero-shot (CoT)		0.7032	**1.0000**	0.8258	0.9975	0.5012	0.0417	0.5012	0.7032
PlagBench few-shot (vanilla)		0.7027	**1.0000**	0.8254	1.0000	0.5000	0.0000	0.5576	0.7280
PlagBench few-shot (CoT)		0.7027	**1.0000**	0.8254	1.0000	0.5000	0.0000	0.5197	0.7111
SCDG (Ours)									
SCDG & Qwen3-8B	Average	0.8914	0.9602	0.9233	0.2748	0.8419	0.7269	0.9161	0.9389
	Top-q	0.9095	0.9696	0.9386	0.2277	0.8708	0.7831	0.9222	0.9379
SCDG & Ministral-3-8B	Average	0.9117	0.9686	0.9392	0.2203	0.8737	0.7863	<u>0.9334</u>	<u>0.9488</u>
	Top-q	0.9124	<u>0.9712</u>	0.9411	0.2215	0.8767	0.7914	0.9236	0.9351
SCDG & Llama-3.1-8B	Average	**0.9183**	0.9696	<u>0.9433</u>	**0.2042**	**0.8835**	<u>0.8008</u>	**0.9387**	**0.9518**
	Top-q	<u>0.9171</u>	0.9707	**0.9449**	<u>0.2067</u>	<u>0.8816</u>	**0.8046**	0.9330	0.9449

Table 1: Pair-level binary classification on PAN25. Entries are per-metric medians over the same 20 hash-locked test partitions. For each backbone, Average aggregates all eligible token-level gains, whereas Top-q aggregates the largest q proportion of gains.

candidates from three fixed first-stage configurations: **equal-weight three-route RRF** (Cormack, Clarke, and Büttcher 2009), and **DAAC-Full**, reporting each before and after reranking.

Evaluation Metrics. For pairwise detection, F_1 is primary; we additionally report precision, recall, FPR, balanced accuracy, MCC, AUROC, and AP. For retrieval, nDCG@10 is primary, supplemented by nDCG@100, Recall@10/100/1000, MRR, and MAP. nDCG preserves the original relevance grades, whereas the remaining retrieval metrics treat relevance > 0 as relevant. For same-topic confounding, we report a proxy false-positive rate by operationally treating same-cluster article pairs as non-reuse cases. Lower values indicate a greater ability to distinguish source reuse from shared topic or event information.

Implementation Details. We instantiate SCDG with three frozen autoregressive backends: Qwen3-8B-Base, Meta-Llama-3.1-8B, and Ministral-3-8B-Base-2512. The unconditional and source-conditioned passes score identical target-token IDs, differing only in whether the candidate source is included in the context. For pairwise detection, aggregation hyperparameters and decision thresholds are selected on the training portion of each split and fixed for test evaluation. For PAN26, SCDG reranks the fixed Top-1,000 candidates from each first-stage system, with all aggregation and fusion hyperparameters selected on development queries only. Full implementation and computing-environment details are reported in the supplement.

Results and Analyses

Pair-Level Binary Classification

Overall performance. Table 1 shows that SCDG provides substantially stronger pairwise discrimination than the lexical, embedding, and prompted-LLM baselines across all three backends. The perfect recall of several PlagBench variants results from predicting nearly every pair as positive, as reflected by their near-unit FPR and near-zero MCC. SCDG instead combines high recall with substantially stronger rejection of non-source pairs. Performance is broadly consistent across Qwen, Ministral, and Llama, with Llama providing the strongest overall point estimates, indicating that the source-conditioned signal is not specific to a single model family.

Effect of sparse aggregation. Top-q aggregation yields slightly higher point-estimate F_1 across all three backends, consistent with source evidence being localized within only part of a suspicious document. When source-derived content occupies only a limited portion of a suspicious document, averaging over all target tokens mixes strong positive gains on reused passages with weak, zero, or negative gains on unrelated content, thereby diluting the document-level dependency signal. Top-q mitigates this effect by emphasizing the upper tail of the token-level gain distribution, where localized source evidence is more likely to concentrate. This interpretation is also consistent with the controlled source-retention results, which show stronger gain responses on plagiarism-labeled content than on newly written content. However, the improvement is metric-dependent: Top-q favors F_1 and MCC, whereas average aggregation retains stronger precision, FPR, AUROC, and AP for the Llama backend. Sparse aggregation therefore changes the operating characteristics of SCDG rather than uniformly improving its discrimination, making the preferred aggregation dependent on the application objective.

Full-Corpus Source Retrieval

Retrieval granularity and fusion. Table 2 shows that retrieval granularity strongly affects candidate quality. Sentence-level dense retrieval substantially outperforms full-

Method	nDCG@10	nDCG@100	R@10	R@100	R@1000	MRR	MAP
Baselines							
BM25-full	0.4962	0.5479	0.5188	0.7425	0.9221	0.7388	0.4175
BM25-sentence	0.5049	0.5629	0.6038	0.8596	0.9821	0.7260	0.4328
BGE-M3 Dense-full	0.4483	0.5022	0.5200	0.7563	0.9221	0.7586	0.4185
BGE-M3 Dense-sentence	0.6240	0.6678	0.7321	0.9096	0.9708	0.8436	0.6071
Four-way CombSUM	0.6764	0.7233	0.7458	0.9342	0.9838	0.9359	0.6578
Linq sentence-to-chunk dense	0.7430	0.7679	0.8442	0.9500	0.9854	0.9468	0.7543
SCDG (our method based on Top-1000)†							
Three-Route RRF Coarse Ranking	0.6864	0.7321	0.7688	0.9600	**0.9900**	0.9454	0.6570
+ fixed SCDG reranking	0.8139	0.8322	0.8742	0.9550	**0.9900**	**0.9925**	0.8332
DAAC coarse ranking	0.7678	0.7892	0.8579	0.9517	0.9854	0.9642	0.7737
+ fixed SCDG reranking	**0.8315**	**0.8433**	**0.9042**	**0.9625**	0.9854	**0.9925**	**0.8532**

Table 2: PAN26 full-corpus source retrieval over all 200 queries against 86,822 candidate sources. Every row is recomputed with the same qrels and evaluator; nDCG uses the original three relevance grades, while recall, MRR, and MAP use relevance > 0. Each indented row applies the same development-selected Qwen3-8B SCDG configuration (Top-2%, $N = 1000$, $\alpha = .75$) to exactly the Top-1000 candidates in the preceding row, so Recall@1000 cannot change within a pair. Boldface and underlining denote the best and second-best distinct values, respectively, in each column; all tied values receive identical formatting.

Method	Proxy FPR (%) ↓
Lexical and retrieval baselines	
BM25 pair score	71.1250%
Max 4-gram containment	31.5625%
4-gram logistic	29.9375%
SCDG statistics and auxiliary calibrators	
Average SCDG	62.8760%
Positive-gain rate SCDG	8.6875%
Gain-distribution logistic	**0.1250%**

Table 3: Proxy false-positive rates on the cluster-disjoint Multi-News same-topic stress set.

document dense retrieval at both the head and deeper cutoffs, whereas sentence segmentation primarily improves the deeper recall of BM25. Linq sentence-to-chunk retrieval is the strongest single-route baseline, and conventional CombSUM does not surpass it at the head of the ranking. This indicates that additional retrieval routes are useful only when their relative reliability is properly controlled.

Effect of SCDG reranking. Across both matched candidate sets, SCDG improves nDCG@10 and MAP without changing Recall@1000, showing that its gains arise from reordering rather than additional retrieval. The larger improvement over RRF suggests that SCDG complements a strong first-stage retriever, with most gains concentrated near the ranking head. This promotion of stronger contributing sources can move weaker relevant candidates across deeper cutoffs, explaining the slight Recall@100 decrease for RRF. As MRR is nearly saturated, nDCG and MAP are more informative for evaluating multi-source ordering. DAAC with SCDG yields the strongest displayed endpoint, whereas the matched RRF comparison more clearly isolates the contribution of reranking.

Robustness to Same-Topic Confounding

Table 3 reports proxy false-positive rates on the Multi-News same-topic stress set, where lower values indicate stronger resistance to topical confounding. BM25, lexical baselines, and Average SCDG remain highly sensitive to shared topic or event information, with false-positive rates of 29.9375–71.1250%. In contrast, Positive-gain-rate SCDG reduces the rate to 8.6875%, while the gain-distribution logistic variant reaches 0.1250%. These results show that topical robustness depends on how the token-level gain distribution is summarized and calibrated, rather than on the mean gain alone. The corresponding PAN25 results in the supplementary materials characterize the trade-off between detection utility and same-topic robustness.

Conclusions

In this work, we introduced SCDG, a directional and training-free framework that measures the incremental predictive evidence supplied by a candidate source through description-length contrast. Controlled interventions support its sensitivity to retained source evidence, while experiments demonstrate strong performance on PAN25 pairwise detection and PAN26 candidate-source ranking across frozen language-model backends. The Multi-News test further indicates robustness to same-topic and same-event confounding. Nevertheless, SCDG is more computationally expensive than conventional similarity measures and has been evaluated mainly on English scientific documents and a constructed news-domain stress set. Future work should improve scoring efficiency, broaden multilingual and cross-domain evaluation, localize passage-level evidence, and study more complex source reuse, while leveraging its token-level decomposition for interpretable evidence tracing. Combined with high-recall retrieval, SCDG enables scalable source attribution without task-specific model updates, supporting more transparent academic-integrity screening and source verification in scholarly writing.

References

Bao, G.; Zhao, Y.; Teng, Z.; Yang, L.; and Zhang, Y. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *The Twelfth International Conference on Learning Representations*.

Bevendorff, J.; Fröbe, M.; Greiner-Petter, A.; Jakoby, A.; Mayerl, M.; Nakov, P.; Plutz, H.; Potthast, M.; Stein, B.; Ta, M. N.; Wang, Y.; and Zangerle, E. 2026. Overview of PAN 2026: Voight-Kampff Generative AI Detection, Text Watermarking, Multi-Author Writing Style Analysis, Generative Plagiarism Detection, and Reasoning Trajectory Detection. arXiv:2602.09147.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335. Bangkok, Thailand: Association for Computational Linguistics.

Chen, J.; Zhu, X.; Liu, T.; Chen, Y.; Chen, X.; Yuan, Y.; Leong, C. T.; Li, Z.; Tang, L.; Zhang, L.; Yan, C.; Mei, G.; Zhang, J.; and Zhang, L. 2025. Imitate Before Detect: Aligning Machine Stylistic Preference for Machine-Revised Text Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22): 23559–23567.

Choi, C.; Kim, J.; Lee, S.; Kwon, J.; Gu, S.; Kim, Y.; Cho, M.; and Sohn, J.-y. 2024. Linq-Embed-Mistral Technical Report. arXiv:2412.03223.

Cormack, G. V.; Clarke, C. L. A.; and Büttcher, S. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, 758–759. Boston, Massachusetts, USA: Association for Computing Machinery. ISBN 978-1-60558-483-6.

Delétang, G.; Ruoss, A.; Duquenne, P.-A.; Catt, E.; Genewein, T.; Mattern, C.; Grau-Moya, J.; Wenliang, L. K.; Aitchison, M.; Orseau, L.; Hutter, M.; and Veness, J. 2024. Language Modeling Is Compression. In *International Conference on Learning Representations*.

Foltýnek, T.; Meuschke, N.; and Gipp, B. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys*, 52(6): 112:1–112:42.

Fox, E. A.; and Shaw, J. A. 1994. Combination of Multiple Searches. In Harman, D. K., ed., *The Second Text REtrieval Conference (TREC-2)*, number 500-215 in NIST Special Publication, 243–252. Gaithersburg, Maryland: National Institute of Standards and Technology.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Greiner-Petter, A.; Fröbe, M.; Wahle, J. P.; Ruas, T.; Gipp, B.; Aizawa, A.; and Potthast, M. 2025. Overview of the Plagiarism Detection Task at PAN 2025. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, 3575–3585. CEUR-WS.org.

Lee, J.; Agrawal, T.; Uchendu, A.; Le, T.; Chen, J.; and Lee, D. 2025. PlagBench: Exploring the Duality of Large Language Models in Plagiarism Generation and Detection. 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)*, 7519–7534. Albuquerque, New Mexico: Association for Computational Linguistics.

Lee, J.; Le, T.; Chen, J.; and Lee, D. 2023. Do Language Models Plagiarize? In *Proceedings of the ACM Web Conference 2023*, 3637–3647. Association for Computing Machinery.

Li, M.; and Vitányi, P. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. New York, NY: Springer, 3 edition. ISBN 978-0-387-33998-6.

Li, Z.; Huang, C.; Wang, X.; Hu, H.; Wyeth, C.; Bu, D.; Yu, Q.; Gao, W.; Liu, X.; and Li, M. 2025. Lossless data compression by large models. *Nature Machine Intelligence*, 7: 794–799.

Liu, A. H.; Khandelwal, K.; Subramanian, S.; Jouault, V.; Rastogi, A.; Sadé, A.; Jeffares, A.; Jiang, A.; Cahill, A.; Gavaudan, A.; et al. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

MacKay, D. J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press. ISBN 978-0-521-64298-9.

Mahoney, M. V. 1999. Text Compression as a Test for Artificial Intelligence. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 970. AAAI Press.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 24950–24962. PMLR.

Potthast, M.; Eiselt, A.; Barrón-Cedeño, A.; Stein, B.; and Rosso, P. 2011. Overview of the 3rd International Competition on Plagiarism Detection. In *Notebook Papers of CLEF 2011 Labs and Workshops*. Amsterdam, The Netherlands.

Potthast, M.; Gollub, T.; Hagen, M.; Graßegger, J.; Kiesel, J.; Michel, M.; Oberländer, A.; Tippmann, M.; Barrón-Cedeño, A.; Gupta, P.; Rosso, P.; and Stein, B. 2012. Overview of the 4th International Competition on Plagiarism Detection. In Forner, P.; Karlgren, J.; and Womser-Hacker, C., eds., *Working Notes for CLEF 2012 Conference*, volume 1178 of *CEUR Workshop Proceedings*. Rome, Italy: CEUR-WS.org.

Potthast, M.; Hagen, M.; Beyer, A.; Busse, M.; Tippmann, M.; Rosso, P.; and Stein, B. 2014. Overview of the 6th International Competition on Plagiarism Detection. In *Working Notes for CLEF 2014 Conference*, volume 1180 of *CEUR Workshop Proceedings*, 845–876. CEUR-WS.org.

Potthast, M.; Hagen, M.; Gollub, T.; Tippmann, M.; Kiesel, J.; Rosso, P.; Stamatatos, E.; and Stein, B. 2013. Overview of the 5th International Competition on Plagiarism Detection.

In *Working Notes for CLEF 2013 Conference*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Potthast, M.; Stein, B.; Barrón-Cedeño, A.; and Rosso, P. 2010. An Evaluation Framework for Plagiarism Detection. In *Coling 2010: Posters*, 997–1005. Beijing, China: Coling 2010 Organizing Committee.

Rissanen, J. 1978. Modeling by Shortest Data Description. *Automatica*, 14(5): 465–471.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27: 379–423, 623–656.

Shannon, C. E. 1951. Prediction and Entropy of Printed English. *The Bell System Technical Journal*, 30(1): 50–64.

Wahle, J. P.; Ruas, T.; Kirstein, F.; and Gipp, B. 2022. How Large Language Models Are Transforming Machine-Paraphrase Plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 952–963. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Wahle, J. P.; Ruas, T.; Meuschke, N.; and Gipp, B. 2021. Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries*, 226–229. IEEE.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, X.; Pan, L.; Zhao, X.; Chen, H.; Petzold, L. R.; Wang, W. Y.; and Cheng, W. 2024. A Survey on Detection of LLMs-Generated Content. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 9786–9805. Miami, Florida, USA: Association for Computational Linguistics.