COMBATING DATA LAUNDERING IN LLM TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data rights owners can detect unauthorized data use in *large language model* (LLM) training by querying with proprietary samples. Often, superior performance (e.g., higher confidence or lower loss) on a sample relative to the untrained data implies it was part of the training corpus, as LLMs tend to perform better on data they have seen during training. However, this detection becomes fragile under *data laundering*, a practice of transforming the stylistic form of proprietary data, while preserving critical information to obfuscate data provenance. When an LLM is trained *exclusively* on such laundered variants, it no longer performs better on originals, erasing the signals that standard detections rely on. We counter this by inferring the unknown *laundering transformation* from black-box access to the *target* LLM and, via an *auxiliary* LLM, synthesizing queries that mimic the laundered data, even if rights owners have only the originals. As the search space of finding true laundering transformations is infinite, we abstract such a process into a high-level transformation *goal* (e.g., "lyrical rewriting") and concrete *details* (e.g., "with vivid imagery"), and introduce *synthesis data reversion* (**SDR**) that instantiates this abstraction. **SDR** first identifies the most probable *goal* that synthesis should step into to narrow the search; it then iteratively refines *details*, such that synthesized queries gradually elicit stronger detection signals from *target* LLM. Evaluated on the MIMIR benchmark against diverse laundering practices and *target* LLM families (Pythia, Llama2, and Falcon), **SDR** consistently strengthens data misuse detection, providing a practical countermeasure to data laundering.

1 INTRODUCTION

Large language models (LLMs) now generate text with human-level fluency and stylistic diversity, driving adoption in medicine (Liu et al., 2025), education (Yan et al., 2024), and other high-stakes applications. Such remarkable capabilities demand training LLMs on large-scale high-quality corpora (Wang et al., 2025), whose collection and use, however, are often constrained by privacy and copyright (Li et al., 2023b). A pressing compliance question is whether a deployed LLM was trained on copyrighted or sensitive material without authorization. In post-hoc *unauthorized training data detections*, a data rights owner queries the *target* LLM with proprietary "candidate" texts and compares a per-sample score, e.g., loss (Zhang et al., 2024) or calibrated confidence proxy (Xie et al., 2024), against the score distribution over a held-out non-training texts corpus, following Carlini et al. (2022). The memorization effect of LLMs (Li et al., 2025) implies that training samples tend to receive lower loss or higher confidence, such that a statistically significant score gap indicates the queried sample likely influences training (Figure 1 (a)); mainstream detection methods perform reliably in this "query with originals" regime (see Table 1 "Orig." columns).

This regime presumes that the *target* LLM is always trained on the rights owner's proprietary texts in their *original form*. In practice, however, natural language is malleable; core information and semantics can still be preserved under extensive stylistic and structural transformations (Barzilay & McKeown, 2001; Bhagat & Hovy, 2013) through human writing, programmatic paraphrase, back-translation (Dolan & Brockett, 2005; Bannard & Callison-Burch, 2005), or more recent LLM-enabled large-scale synthesis (Witteveen & Andrews, 2019; Liu et al., 2024b). When an LLM is trained *exclusively* on such transformed surrogates, it does *not* memorize original data and, no longer exhibits a reliable performance gap when queried with the originals (Figure 1 (b)), which erases the signals that unauthorized data detections rely on. Empirically, consider a target Llama-2 (Touvron et al., 2023) trained on a corpus stylistically transformed from Wikipedia articles into lyrics, we find

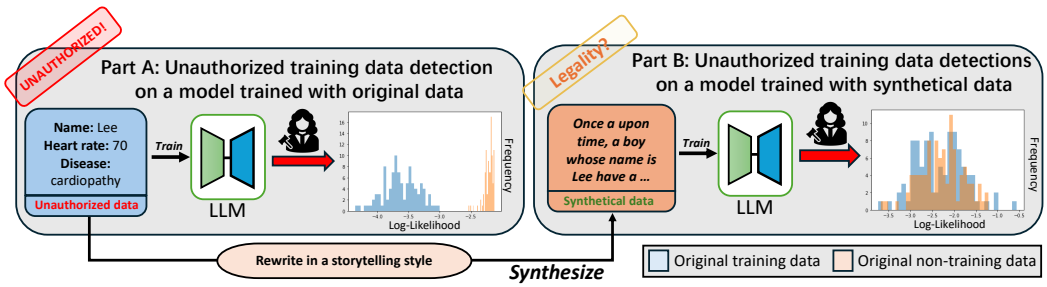

Figure 1: Illustration of how data laundering undermines existing unauthorized training data detections. When unauthorized data is directly used for training, LLMs tend to memorize the unauthorized training data. Training samples exhibit lower loss than non-training data, as shown in Part A. The log-likelihood distributions of training and non-training samples diverge clearly, enabling identification. However, when trained on laundered unauthorized data; as shown in Part B, the distributions of the unauthorized data and non-training samples no longer diverge, preventing reliable identification.

that mainstream unauthorized data detection methods, when tested on the originals (i.e., Wikipedia articles), perform no better than random chance (Table 1 "Syn." columns).

This fragility enables *data laundering*: deliberate obfuscation of data provenance through semantic-preserving transformation to conceal large-scale unauthorized use. Model providers can transform entire proprietary corpora, e.g., synthetically alter personal health records into children's-story-style narratives that retain substantive content, then train on the derivatives, and assert that the original records never entered the training, although sensitive information may still influence the model and can be further leaked. In realistic detections, data rights owners typically have *black-box* access to deployed *target* LLMs via an inference-only API; providers rarely disclose their preprocessing or sourcing pipelines, nor training artifacts. This opacity, where the potential laundering procedure is *unknown* to rights owners, creates a risk to intellectual property and privacy, motivating our question: *How can a data rights owner detect unauthorized data use from black-box model access when the data has been laundered through an unknown transformation?*

Effective detection requires queries that stylistically and structurally resemble what the *target* LLM observed during training. When the laundering transformation is unknown, crafting effective queries becomes an unbounded search over possible data alterations. Our key idea here is to shift the problem from locating specific laundered samples to inferring the *laundering transformation* itself (formalized in Section 3). We make this problem tractable by modeling the unknown transformation as a synthesis procedure defined by a two-level *goal-details* abstraction. A high-level *goal* that captures the primary language register shift[1] (e.g., "rewrite into lyrics") and concrete *details* that refine stylistic and formatting constraints further (e.g., imagery, voice, and rhyme density). Leveraging modern LLMs as controllable generators (Liang et al., 2024a), we instantiate this abstraction as a language-*prompted* specification executed by an *auxiliary* LLM to synthesize candidate surrogates under explicit controls. This *goal-details* schema is compatible with commonly used prompt templates (Mao et al., 2025), allows *goal* to set the coarse-grained stylistic and structural transformation, while *details* provide fine-grained, data-driven synthesis refinement[2].

In Section 4, we introduce *synthesis data reversion* (**SDR**), a two-stage search that returns (i) a *goal-details* synthesis specification and (ii) a set of "training-like" queries synthesized from proprietary texts under detection (i.e., candidate set) and compared against a reference non-training texts (i.e., held-out set), enabling off-the-shelf detection methods, e.g., (Xie et al., 2024) on a *target* LLM.

Mirroring the schema, stage 1 determines the most likely laundering *goal* by screening an established taxonomy of 23 registers (Myntti et al., 2025). For each register, we pre-define a standard rewriting prompt and, with the *auxiliary* LLM, produce short rewrites of a small seed of candidate texts; from these samples we extract a common opening template that captures how the register typically begins. We then task the *target* LLM to score register-conditioned rewrites and keep the few registers that

[1] A register is a situational variety of language shaped by purpose, audience, and medium; examples include news, academic prose, instructions, and lyrics (Agha, 2004).

[2] We do not claim true laundering always follows this schema; it is adopted only as a search strategy for synthesizing queries in "training-like" style, which will be used for detection.

Table 1: Performance of unauthorized training data detection on Llama-2 (Touvron et al., 2023) models fine-tuned with either the original MIMIR-wiki (Deng et al., 2023) dataset (Orig.) or its laundered version (Syn.) generated by GPT-4o (Hurst et al., 2024) using the prompt "rewrite in a lyrical style, ensuring the imagery is vivid". Evaluation metrics are defined in Section 5.

Methods	AUC		ASR		TPR@5%	
	Orig.	Syn.	Orig.	Syn.	Orig.	Syn.
Loss (Yeom et al., 2018)	1.000	0.539	1.000	0.565	1.000	0.040
Ref (Carlini et al., 2022)	0.971	0.603	0.920	0.610	0.850	0.100
Zlib (Carlini et al., 2021)	1.000	0.521	1.000	0.535	1.000	0.080
Min-K (Shi et al., 2023)	1.000	0.563	1.000	0.575	1.000	0.040
Recall(Xie et al., 2024)	0.999	0.558	0.995	0.565	1.000	0.000

best match the *target* LLM's preferences. Lastly, for each shortlisted register, we synthesize full rewrites of the candidate and reference texts, run unauthorized-use detection, and select the register that maximizes the detection metrics (Algorithm 1). This yields the initial *goal* specification that stage 2 will refine with *details*. Starting from the selected *goal* and its standard prompt, in stage 2 we iteratively infer the missing fine-grained details that make rewrites resemble what the *target* LLM likely saw during training. In each iteration, we sample a seed of proprietary texts; the *auxiliary* LLM rewrites each under the current specification, and *target* LLM generates the next span following the rewrite's opening sentence. The *auxiliary* LLM then summarizes the differences between pairs of rewrites and target-generated follow-ons into refinements to the current specification. Upon using the revised specification to synthesize rewrites for both candidate and reference sets, we accept the revision only if it improves the unauthorized-use detection performance. The loop repeats until the gains plateau or the maximum iteration is reached, yielding a *goal-details* specification and "training-like" surrogates usable with off-the-shelf detections under laundering (Algorithm 2).

In Section 5, we evaluate **SDR** on MIMIR benchmark (Deng et al., 2023) across *target* LLM families (Pythia (Biderman et al., 2023), Llama-2 (Touvron et al., 2023), falcon (Zhang et al., 2022)), and *auxiliary*-LLM choices (DeepSeek (Liu et al., 2024a), GPT-4o (Roumeliotis & Tselikas, 2023), Claude (Wu et al., 2023)) under diverse simulated large-scale laundering procedures. **SDR** consistently strengthens off-the-shelf standard detection methods in *all* detection metrics with ablation studies showing that both stages contribute to the gains.

Contributions. This study presents (i) a data laundering-aware, post-hoc unauthorized data detection formulation for black-box LLMs. (ii) a *goal-details* abstraction that constructs a tractable search space over undisclosed laundering transformation. (iii) **SDR**, a practical two-stage method that restores the effectiveness of standard detection methods even under laundering. Together, we hope this study establishes an actionable blueprint for data rights holders to verify unauthorized training under data laundering in black-box LLMs and *raises practitioners' awareness of data laundering*.

2 Unauthorized Data Detection

Verifying the provenance of data used to train LLMs is a cornerstone of trustworthy AI, with critical implications for copyright compliance, data privacy, and license enforcement (Li et al., 2023a). The field has developed two main strategies for data governance, including proactive measures applied before/during training and post-hoc detection of trained models.

Proactive defenses are approaches that prevent or trace data misuse from the outset. *Data watermarking* embeds imperceptible signals, such as stylistic patterns, directly into training data, which can subsequently be detected in a model's outputs to establish provenance (Liang et al., 2024b). Similarly, *parameter watermarking* (Kirchenbauer et al., 2023) embeds ownership signals within the model's weights. They necessitate that data owners anticipate misuse and modify the data before its collection or training, making them inapplicable for auditing pre-trained models. Differential privacy (Dwork, 2008), on the other hand, offers cryptographic-style guarantees against memorization by introducing calibrated noise during the training process, but frequently incurs a substantial penalty on model utility; it is rarely adopted in training LLMs where model performance is paramount. In addition, dataset documentation frameworks like *datasheets for datasets* (Gebru et al., 2021) and *model cards* (Mitchell et al., 2019) foster transparency regarding training data composition. These are, however, voluntary disclosures and cannot verify the absence of undisclosed data sources, nor can they audit existing models whose provenance may be deliberately obscured. In

brief, proactive defenses are essential but have inherent limitations; they are not universally implemented, and do not provide a mechanism for auditing existing models trained without such foresight.

Post-Hoc detection, in contrast, seeks to determine if specific data was used to train a deployed model *after* its development, often with black-box access. This task is often instantiated by techniques derived from the *membership inference* literature (Shokri et al., 2017). A post-hoc detector aims to distinguish a model's training data (i.e., members) from unseen data (i.e., non-members) by exploiting statistical differences in model behavior (Li et al., 2025). Since overparameterized models, e.g., neural nets, have shown a strong tendency to memorize their training data, they frequently demonstrate higher confidence or lower loss on member samples compared to non-members (Carlini et al., 2021; 2022). This performance difference constitutes the primary signal that post-hoc detection methods are designed to exploit. Adapting post-hoc detectors to modern LLMs presents challenges due to prohibitive costs in training shadow models (Carlini et al., 2022). Thus, cutting-edge detectors for LLMs typically analyze target model's intrinsic signals (i.e., target model's own output) directly, encompassing loss-based signals (Ye et al., 2024), likelihood-based comparisons (Shi et al., 2023; Zhang et al., 2024). Xie et al. (2024) leveraged calibrated confidence scores to construct robust tests for membership. The practical objective of these techniques is broader than membership inference: to provide data rights owners with reliable means for detecting unauthorized use. Empirically, querying the target model with proprietary data, mainstream methods are effective across commonly used benchmarks.

Data laundering breaks post-hoc detection. It is worth noting that, all existing post-hoc detection methods are designed and evaluated on the "query with originals" regime, where the rights holder queries original proprietary texts to the *target* LLM (in the context of natural language) and compares intrinsic signal-based scores against the non-member corpus. This regime ignores that language exhibits pliability in form and usage, which creates a blind spot when *target* LLM was trained on surrogates—semantics-preserving but stylistically or structurally altered variants produced by paraphrase and back-translation (Barzilay & McKeown, 2001; Bannard & Callison-Burch, 2005), register/style transfer (e.g., news $\rightarrow$ instructions), or large-scale LLM-based rewriting (Witteveen & Andrews, 2019; Zeleke et al., 2025)—rather than on originals. This mechanism, when exploited by model providers, enables practices to evade unauthorized-use detections, which we call *data laundering*. Since *target* LLM never truly saw originals, its memorization effect is tied to the surrogates; members no longer enjoy systematic intrinsic score advantages on the originals, collapsing the intrinsic score gap between members and non-members and failing data rights owners who apply standard unauthorized-use detectors to *target* LLM with originals.

Can post-hoc detection be restored? As aforementioned, the threat is compounded by the opacity of real-world LLM deployment. Typically, data rights owners have only black-box access to *target* LLM, and model providers rarely disclose training details. The specific laundering transformation, if one was used at all, is invisible to rights owners. The space of potential transformations is infinite, making brute-force search intractable. This gives rise to a practical impossibility for auditors: without knowing the hidden transformation, one cannot hope to produce an *exact* training-time query that would elicit the memorization signal from *target* LLM. Thus, instead of building a new detector robust to laundered samples, we reframe the problem and turn to inferring the *laundering process* itself. By reverse-engineering the transformation's properties from the black-box *target* LLM's behavior, we can synthesize queries that are "training-like" to restore the statistical signals needed for detection. This approach, if successful, allows us to re-enable standard, off-the-shelf detectors, making them effective even in the presence of data laundering.

3 Reverse-engineering the Laundering Transformation

As established, directly finding the exact laundered data is intractable. Our approach, therefore, is to find a generative/synthesis process that *mimics* the unknown laundering transformation, leveraging a powerful, auxiliary LLM as a controllable *transformation simulator* under prompt specification. Specifically, we search for a natural language prompt that, when given to *auxiliary* LLM, causes it to produce outputs that stylistically mimic the data *target* LLM was likely trained on.

Objective. Given an off-the-shelf detector, a *target* LLM, an *auxiliary* LLM, a candidate set of originals (suspected members), and a held-out set (known non-members), we aim to find an estimated transformation, specified by prompted-based synthesis with *auxiliary* LLM, that maximizes

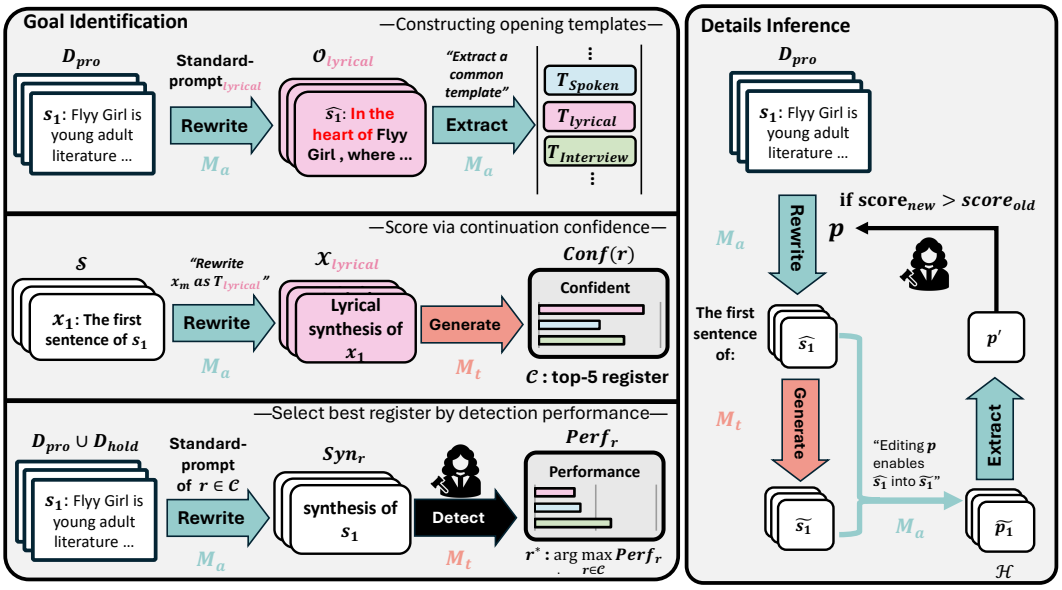

Figure 2: Pipeline of the **SDR** framework. In the **goal identification stage** (left part), **SDR** tries to find the register that is closely aligned with the laundering goal (See Algorithm. 1). The **details inference stage** (right part) tries to infer the remaining details in the laundering process (See Algorithm. 2).

the expected detection score separation between the two sets. Upon obtaining this transformation, we generate new "training-like" queries using *auxiliary* LLM, and use them with any detectors.

Managing the prompt space. Still, the search space of possible natural language prompts is effectively infinite (Zhang et al., 2025), making an unconstrained search infeasible. To solve this, we introduce a structured abstraction that makes the search tractable, leveraging established principles from prompt engineering (Mao et al., 2025) and linguistic theory (Agha, 2004; Myntti et al., 2025).

We first structure the estimated transformation prompts using a *goal-details* schema, motivated by recent work showing that effective prompts can be decomposed into a core directive and supplementary modifiers such as context, constraints, or output format (Mao et al., 2025). We adapt this structure as follows: The *goal* defines the transformation's high-level intent and dominant stylistic shift (e.g., "rewrite in a lyrical style"); the *details* aggregate all other components that refine the output (e.g., "ensuring the imagery is vivid"). To further reduce the search space for *goals*, we ground it in an established linguistic taxonomy of 23 registers that collectively cover primary communicative forms (Henriksson et al., 2024). Our task now becomes selecting the most probable *goal* from this finite set of registers[3] and refining *details* within it. By combining the LLM-based transformation simulator with this structured *goal-details* prompt abstraction, we transform the intractable optimization problem into a constrained search. This leads to the two-stage method detailed next.

4 Synthesis Data Reversion

We propose a two-stage framework, *synthesis data reversion* (**SDR**), to reverse the laundered data used to train the target model. The first stage, the *goal identification* stage, aims to infer the laundering *goal*. The second stage, *details inference*, aims to recover the supplementary conditions of the laundering process. Figure 2 overviews the pipeline, and a detailed description is in Appendix D.

Goal identification stage. At this stage, our method tries to find the register in the established 23 registers that is closely aligned with the laundering *goal* (details are in Algorithm 1). A straightforward approach would be to synthesize all proprietary samples into each register using the auxiliary LLM and to determine which register synthesis improves the performance of unauthorized retaining data detection mostly. However, extensively querying the auxiliary LLM with long-token sequences

[3]We acknowledge that this taxonomy was not designed for data laundering and thus has limitations, which we discuss in Appendix I.

Algorithm 1 Goal identification stage

Require: Proprietary originals $\mathcal{D}_{\text{pro}}$, held-out data $\mathcal{D}_{\text{held}}$, target model M_t, auxiliary LLM M_a, GPT-5, set of 23 registers $\mathcal{R}$, sample size n and m
Ensure: the register $r^* \in R$ that is closely aligned with the laundering directive.

—Constructing opening templates—
for all $r \in \mathcal{R}$ **do**
  Standard-prompt$_r \leftarrow$ GPT-5("Give me a prompt that can transfer text into register r")
  $\mathcal{O}_r \leftarrow \{$ The first sentence of M_a(Standard-Prompt$_r, s) \mid s \in$ UniformSample$(\mathcal{D}_{\text{pro}}, n)\}$
  $\text{T}_r \leftarrow M_a$("Extract a common template.", $\mathcal{O}_r$)
end for
—Score via continuation confidence—
for all $r \in \mathcal{R}$ **do**
  $\mathcal{S} \leftarrow \{$ The first sentence of $s) \mid s \in$ UniformSample$(\mathcal{D}_{\text{pro}}, m)\}$
  $\mathcal{X}_r \leftarrow \{ M_a$("Rewrite x as T_r") $\mid x \in \mathcal{S}\}$
  for $j \leftarrow 1$ **to** m **do**
    $c_j \leftarrow$ Average next token confidence of $M_t(\mathcal{X}_r[j])$
  end for
  $\text{Conf}(r) \leftarrow \frac{1}{m}\sum_j c_j$
end for
$\mathcal{C} \leftarrow$ top-5 registers with largest $\text{Conf}(r)$
—Select best register by detection performance—
for all $r \in \mathcal{C}$ **do**
  $\text{Syn}_{\text{r}} \leftarrow \{ M_a$(Standard-prompt$_r, d) \mid d \in \mathcal{D}_{\text{pro}} \cup \mathcal{D}_{\text{held}}\}$
  $\text{Perf}_r \leftarrow$ Unauthorized training data detection on M_t using Syn_{r}
end for
$r^* \leftarrow \arg\max_{r \in \mathcal{C}} \text{perf}_r$
return r^*, Standard-prompt$_{r^*}$

for synthesis is costly. To reduce this cost, we rewrite only the opening sentence of each sample into different registers by using the extracted register-specific opening templates. If the proprietary data was laundered under a goal close to one of the registers, the laundered samples will likely have an opening template resembling that register's template. The target model will generate continuations more confidently, conditioning on a familiar opening sentence (Yeom et al., 2018). Thus, by observing which register's opening sentences lead to higher confidence in continuation generation, we can identify at low cost the register closer to the laundering goal.

Specifically, for each register r, we first use GPT-5 (Leon, 2025) [4] to generate a *Standard-prompt* that can synthesize data into that register. Using Standard-prompt, the auxiliary LLM rewrites n samples of proprietary data and abstracts the first sentences of them into an opening template (see constructing opening templates). Following the *Score via continuation confidence*, these templates are applied to rewrite the first sentence of the original data. Rewritten sentences are then provided to the target model and measure the average model's continuation confidence $\text{Conf}(r)$ (Details of the confidence calculation are shown in Appendix B). The top-5 registers with the highest $\text{Conf}(r)$ are retained as possible registers $\mathcal{C}$. Finally, the closest register r^* is selected from $\mathcal{C}$ based on its unauthorized training data detection performance (see "Select best register by detection performance").

Details inference stage. Once the closest register aligned with the laundering directive has been identified, we can reverse the laundered data by synthesizing the original samples into that register. However, this reversion may still diverge from the true laundered training data, as additional *details* may have been applied in the laundering process. The second stage seeks to recover these supplementary details (see Algorithm 2). Directly comparing the closest register synthesis with the true training data would reveal such details, but this is infeasible for the data rights owner unless similar data can be found. In the previous stage, we know that the first sentence of the closest register synthesis resembles that of the target model's training data. Providing such a familiar opening sentence to the target model activates its memory of training data, enabling it to reproduce the corresponding memorized continuations that are similar to the training data. As a result, analyzing the differences between the closest register synthesis and the reproduced continuations enables us to recover the supplementary conditions.

[4] We evaluate the transferability of **SDR** across various scenarios using the *Standard-prompts* generated by GPT-5 (Section 5), and the results consistently show strong efficiency and robustness.

Algorithm 2 Details inference stage

Require: register r^* and Standard-prompt$_{r^*}$ got from Algorithm 1 , proprietary data D_{pro}, held-out data $\mathcal{D}_{\text{held}}$, target model M_t, auxiliary LLM M_a, iteration budget K, sample size l
Ensure: Reversed prompt

$p =$ Standard-prompt$_{r^*}$
function CONDITIONINFERENCE($\mathcal{D}_{\text{pro}}, p, M_t, M_a$)
  for all $s \in$ UnifiedSample($\mathcal{D}_{\text{pro}}, l$) **do**
    $\hat{s} \leftarrow M_a(p, s)$, $\tilde{s} \leftarrow M_t$(the first sentence of $\hat{s}$)
    $\mathcal{H}$.APPEND(M_a("Editing p enables the transofrmation of $\hat{s}$ into $\tilde{s}$"))
  end for
  return M_a("Extract a common prompt.", $\mathcal{H}$)
end function
function EVALUATE($D_{pro}, D_{\text{held}}, p, M_t, M_a$)
  $\text{Syn}_{\text{p}} \leftarrow \{ M_a(p, x) \mid x \in \mathcal{D}_{\text{pro}} \cup \mathcal{D}_{\text{held}} \}$
  $\text{Perf}_r \leftarrow$ Unauthorized training data detection on M_t using Syn_{p}
  return Perf_p
end function
for $k \leftarrow 0$ **to** $K - 1$ **do**
  $p' \leftarrow$ CONDITIONINFERENCE($\mathcal{D}_{\text{pro}}, p, M_t, M_a$)
  $\text{score}_{\text{new}} \leftarrow$ EVALUATE($\mathcal{D}_{\text{pro}}, \mathcal{D}_{\text{held}}, p', M_t, M_a$), $\text{score}_{\text{old}} \leftarrow$ EVALUATE($\mathcal{D}_{\text{pro}}, \mathcal{D}_{\text{held}}, p, M_t, M_a$)
  if $\text{score}_{\text{new}} > \text{score}_{\text{old}}$ **then**
    $p \leftarrow p'$
  end if
end for
return p

Particularly, we first synthesize proprietary samples with the auxiliary LLM using an initial prompt p that can synthesize data into the closest register (the Standard-prompt$_{r^*}$ got from the previous stage). Using the function CONDITIONINFERENCE, the first sentence of the synthesis is fed into the target model to generate continuations. Both the synthesized data and the generated continuations are provided to the auxiliary LLM, which infers the details involved in the laundering process. To determine whether refinement improves performance, we apply the EVALUATE function; if so, the refined prompt replaces the initial one and the process continues iteratively. Through iterative updates, a recovered prompt with refined details is created that best approximates the laundering process, enhancing detection performance on its reversed data.

5 EXPERIMENTS AND RESULTS

Dataset and victim models. We utilize the MIMIR (Deng et al., 2023) benchmark dataset, a widely recognized resource in research on unauthorized training data detection. To evaluate the generality of our method, we select three subsets from MIMIR: Wikipedia, C4, and HackerNews, corresponding to encyclopedia articles, web text, and news reports, respectively. As victim models, we employ several different architectures, including Pythia (Biderman et al., 2023), Falcon (Zhang et al., 2022), and LLaMA-2 (Touvron et al., 2023) to evaluate the robustness of **SDR** across architecture.

Baselines and metrics. We involve five baseline unauthorized training data detections in our experiments: Loss (Yeom et al., 2018), which uses likelihood loss as the membership score; Ref (Carlini et al., 2022), which calibrates input loss via a reference model; Zlib (Carlini et al., 2021), which compresses input loss through entropy coding; Min-K% (Shi et al., 2023) and RecaLL (Xie et al., 2024) as introduced in Section 2. Following prior work in (Carlini et al., 2022), we report three metrics: Area Under the Curve (AUC), Attack success rate (ASR), and True Positive Rate at 5% False Positive Rate (TPR@5%). Details explaining the metrics are shown in Appendix E.

Evaluation settings and implementation details. We evaluate the effectiveness of **SDR** by examining whether it reverses data to enhance the performance of existing unauthorized training data detection methods against target LLMs trained on laundered data. (See Appendix F for details.)

Synthesized prompt setting. We consider two types of prompts that may be applied by the model provider: inside-register and outside-register. Inside-register prompts assume that the model provider synthesizes the original data into one of the 23 sub-registers. For each register, we use GPT-

Table 2: The average performance of each unauthorized training data detection method across data synthesized from **different inside and outside luandering process**. The experience is located on Pythia-6.9B (Biderman et al., 2023), fine-tuned on Wikipedia synthesis. The specific results for each prompt are provided in Appendix H.

	Inside registers			Outside registers		
Method	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%
Recall.	64.7%	63.4%	8.9%	61.7%	61.4%	5.6%
Recall+**SDR**	**76.2%**	**72.0%**	**25.3%**	**73.4%**	**73.3%**	**23.2%**
Loss.	63.7%	62.8%	10.7%	62.7%	62.6%	9.2%
Loss+**SDR**	**76.6%**	**72.6%**	**26.2%**	**75.5%**	**75.5%**	**22.9%**
Ref	68.6%	67.0%	15.0%	67.6%	65.6%	13.2%
Ref+**SDR**	**74.8%**	**70.8%**	**29.9%**	**72.0%**	**72.1%**	**24.2%**
Zlib	63.9%	63.5%	15.2%	63.6%	63.5%	13.9%
Zlib+**SDR**	**68.9%**	**66.7%**	**18.8%**	**68.4%**	**68.4%**	**16.2%**
Min-K	63.5%	62.6%	11.8%	64.2%	62.5%	10.5%
Min-K+**SDR**	**75.1%**	**71.6%**	**25.1%**	**73.6%**	**73.5%**	**22.7%**

Table 3: Comparison of average performance of unauthorized training data detection across **different datasets** trained with synthesis using outside register prompts.

	Wikipedia			Hackernews			C4		
okMethod	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%
Recall	61.7%	61.4%	5.6%	53.9%	56.2%	9.9%	52.2%	55.0%	5.0%
Recall+**SDR**	**73.4%**	**73.3%**	**23.2%**	**61.6%**	**60.3%**	**10.7%**	**65.2%**	**63.0%**	**12.0%**
Loss	62.7%	62.6%	9.2%	54.2%	56.0%	6.3%	58.4%	59.1%	7.2%
Loss+**SDR**	**75.5%**	**75.5%**	**22.9%**	**62.7%**	**59.6%**	**8.7%**	**67.3%**	**64.2%**	**13.0%**
Min-K	64.2%	62.5%	10.5%	53.8%	55.9%	5.7%	57.6%	59.2%	6.2%
Min-K+**SDR**	**73.6%**	**73.5%**	**22.7%**	**61.7%**	**60.8%**	**8.3%**	**66.8%**	**64.7%**	**11.8%**

5 (Leon, 2025) to generate a corresponding prompt. Outside-register prompts are those generated by GPT-5 that do not align with any established register. The full list of inside- and outside-register prompts is provided in Appendix G.

Result analysis across different synthesized prompts. To evaluate whether **SDR** can reverse synthesize training data from different prompts, we use GPT-4o (Roumeliotis & Tselikas, 2023) to synthesize the MIMIR-Wikipedia data into new data with different inside- and outside prompts (mentioned in Section 5) and fine-tune a Pythia-6.9B (Biderman et al., 2023) model. Table 2 shows that across both inside- and outside-register prompts, **SDR** consistently enhances the average performance of the detection. Specifically, the average detection AUC of Loss increases by 12.9% across the inside prompts and 12.8% across the outside prompts. The specific results for each prompt are provided in Appendix H.

Result analysis across different datasets. To evaluate the robustness of **SDR** across different datasets, we applied unauthorized training data detections to Pythia-6.9B models, which were trained with synthesized data from various datasets (Wikipedia, Hackernews, and C4) under outside prompts. Table 3 shows that across all three datasets, **SDR** consistently improves detection performance. For example, Recall achieves a clear AUC gain on all datasets corresponding to an average improvement of 10.8%.

Result analysis across different LLM structures. To evaluate the robustness of **SDR** across different trained model architectures, Table 4 shows the performance of unauthorized training data detections on three different model architectures (Pythia-6.9B, Falcon-7B, and LLaMA-2-7B) fine-tuned with MIMIR-Wikipedia synthesis using outside register prompts. Across all three models, **SDR** consistently enhances detection effectiveness. For example, Recall achieves substantial AUC gains on all models, with an average improvement of 9.3%.

Result analysis with different auxiliary models. We examine a scenario in which the auxiliary LLM applied by the data rights holder for **SDR** differs from the one employed by the model provider used to launder data. As shown in Table 5, in this experiment, we consider that the model provider synthesizes data using GPT-4o, while **SDR** employs auxiliary LLMs such as Claude and DeepSeek for reverse synthesis. Results are averaged using the first ten inside prompts. From the result, we can find that **SDR** achieves an average AUC improvement of 13.5% on GPT-4o, 11.5% on Claude,

Table 4: Comparison of average performance of unauthorized training data detection across three **different model architectures** fine-tuned with synthesis generated by outside register prompts.

	Pythia-6.9B			Falcon-7B			LLaMA-2-7B		
Method	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%
Recall	61.7%	61.4%	5.6%	62.1%	61.8%	8.2%	61.6%	61.6%	8.2%
Recall+**SDR**	**73.4%**	**73.3%**	**23.2%**	**72.4%**	**69.1%**	**23.0%**	**67.5%**	**66.2%**	**12.5%**
Loss	62.7%	62.6%	9.2%	64.4%	64.4%	11.5%	64.5%	64.4%	11.1%
Loss+**SDR**	**75.5%**	**75.5%**	**22.9%**	**71.2%**	**68.0%**	**20.2%**	**73.6%**	**70.4%**	**26.9%**
Min-K	64.2%	62.5%	10.5%	63.3%	62.6%	10.5%	62.5%	62.5%	13.9%
Min-K+**SDR**	**73.6%**	**73.5%**	**22.7%**	**70.2%**	**68.0%**	**20.0%**	**70.9%**	**68.2%**	**22.9%**

Table 5: Comparison of the average performance of unauthorized training data detection with **SDR** using **different auxiliary LLMs**.

	GPT-4o			Claude			DeepSeek		
Method	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%	AUC	ACC	TPR@5%
Recall	64.5%	63.5%	9.9%	66.8%	65.9%	9.6%	65.5%	64.8%	7.4%
Recall+**SDR**	**79.7%**	**75.0%**	**31.6%**	**79.8%**	**75.5%**	**31.2%**	**79.8%**	**75.5%**	**31.2%**
Loss	63.4%	62.2%	14.3%	67.3%	65.3%	12.1%	65.9%	64.7%	12.2%
Loss+**SDR**	**80.3%**	**75.6%**	**32.6%**	**79.1%**	**74.7%**	**32.8%**	**79.0%**	**74.7%**	**32.8%**
Min-K	63.7%	62.3%	11.0%	68.0%	65.5%	14.5%	67.0%	64.6%	14.2%
Min-K+**SDR**	**72.0%**	**75.1%**	**32.8%**	**77.8%**	**73.1%**	**30.3%**	**77.8%**	**73.1%**	**30.3%**

and 12.7% on DeepSeek. These increasing values are close, indicating that **SDR**'s effectiveness is stable across different auxiliary LLMs.

Ablation study. To assess the contribution of each stage in the proposed **SDR** framework, we conduct an ablation study by selectively removing individual stages. By comparing the performance of unauthorized training data detections on the data reversed with the full **SDR** framework against that with individual stages, we can evaluate the necessity of each stage. As shown in Figure 3, directly applying the identified directive from the *goal identification stage* (w/o stage 2) to reverse the synthesized data leads to degradations in all average detection metrics across different inside prompts. In particular, with TPR@5% dropping by 7.5% compared to the full **SDR** framework (SDR). Skipping *the register identification stage* and only relying on the *details inference* (w/o stage 1) causes even more severe degradation, reducing TPR@5% by 12.5%. These results demonstrate that both stages are indispensable.

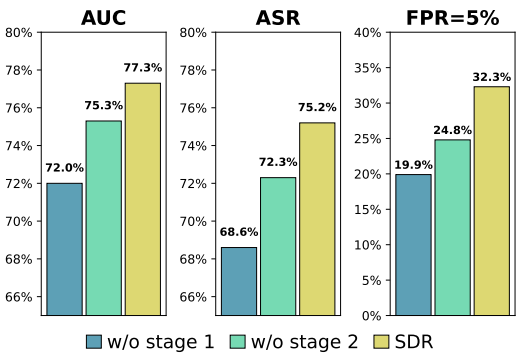

Figure 3: Ablation study on the effectiveness of each stage in **SDR** . Results are reported as the average performance of unauthorized training data detections across different inside prompts. Removing the directive identification stage (w/o stage 1) or the detailed prompt condition inference stage (w/o stage 2) leads to noticeable degradation, while the full SDR consistently achieves the best performance.

6 Conclusion

This paper identified a critical vulnerability in current auditing practices: conventional unauthorized training data detections fail under data laundering, leaving a loophole that enables model providers to obscure the provenance of training data. To address this challenge, we proposed **SDR**, a two-stage framework that reconstructs the synthesis process by inferring a prompt to recover laundered data, thereby restoring the detectability of unauthorized usage. Through extensive evaluation across datasets, model architectures, and auxiliary LLM models, we demonstrated that **SDR** enhances the effectiveness of unauthorized training data detection. In future work, the focus should be on developing finer-grained, task-specific directive taxonomies to improve the accuracy and robustness of prompt reversal. In sum, we believe **SDR** opens a promising direction for developing robust privacy auditing tools against data laundering.

7 Ethics Statement

This work focuses on defending against unauthorized data laundering in LLM training by restoring the effectiveness of post-hoc detection. All experiments were conducted on publicly available datasets, and no proprietary or personal data was used. While there is a risk that our techniques could be misused to improve laundering attacks, our contributions are explicitly framed for defensive purposes, aiming to strengthen accountability and responsible governance in AI systems.

8 Reproducibility Statement

We have made significant efforts to ensure the reproducibility of our results. All datasets used in this paper are publicly available, and details of data synthesis procedures are provided in Appendix G. Complete descriptions of model architectures, training settings, and evaluation protocols are included in the main paper and Appendix F. For each experiment, we specify hyperparameters, implementation details, and the auxiliary LLM prompts used for synthesis and reverse synthesis. Our codebase, built on PyTorch and Hugging Face Transformers, was included in the submitted supplementary material and will be released upon publication to facilitate full replication of our experiments.

References

Asif Agha. Registers of language. *A companion to linguistic anthropology*, 2004.

Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *ACL*, 2005.

Regina Barzilay and Kathleen McKeown. Extracting paraphrases from a parallel corpus. In *ACL*, 2001.

Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational linguistics*, 2013.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *SP*, 2022.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*, 2023.

Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *IWP*, 2005.

Cynthia Dwork. Differential privacy: A survey of results. In *TAMC*, pp. 1–19, 2008.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

Erik Henriksson, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson, and Veronika Laippala. Automatic register identification for the open web using multilingual deep learning. *arXiv preprint arXiv:2406.19892*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *ICML*, 2023.

Maikel Leon. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, 2025.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 2023a.

Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, et al. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*, 2023b.

Muxing Li, Zesheng Ye, Yixuan Li, Andy Song, Guangquan Zhang, and Feng Liu. Membership inference attack should move on to distributional statistics for distilled generative models. *arXiv preprint arXiv:2502.02970*, 2025.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024a.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S Yu. Watermarking techniques for large language models: A survey. *arXiv preprint arXiv:2409.00089*, 2024b.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. Monotonic paraphrasing improves generalization of language model prompting. In *EMNLP Findings*, 2024b.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 2025.

Yuetian Mao, Junjie He, and Chunyang Chen. From prompts to templates: A systematic prompt template analysis for real-world llmapps. In *FSE*, 2025.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *CFAT*, 2019.

Amanda Myntti, Erik Henriksson, Veronika Laippala, and Sampo Pyysalo. Register always matters: Analysis of llm pretraining data through the lens of language variation. *arXiv preprint arXiv:2504.01542*, 2025.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *SP*, 2017.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025.

Sam Witteveen and Martin Andrews. Paraphrasing with large language models. In *ACL Workshop on Neural Generation and Translation*, 2019.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*, 2023.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. *arXiv preprint arXiv:2406.15968*, 2024.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 2024.

Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. Data contamination calibration for black-box llms. *arXiv preprint arXiv:2405.11930*, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*, 2018.

Brook Zeleke, Amish Soni, and Lydia Manikonda. Human or genai? characterizing the linguistic differences between human-written and llm-generated text. In *ACM Web Science Conference*, 2025.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xiang Zhang, Juntai Cao, Jiaqi Wei, Chenyu You, and Dujian Ding. Why prompt design matters and works: A complexity analysis of prompt search space in llms. *arXiv preprint arXiv:2503.10084*, 2025.

A Register Taxonomy

The register taxonomy proposed by Henriksson et al. (2024) defines 23 sub-registers across nine major categories, ranging from narrative forms (e.g., news and sports reports, blogs) and informational texts (e.g., encyclopedia entries, research articles, legal documents) to opinion pieces, persuasive descriptions, interactive discussions (e.g., FAQs, interviews), instructional texts (e.g., recipes), spoken and lyrical registers, and machine-translated content (details are shown in Table 6). This taxonomy provides near-comprehensive coverage of open-domain texts, offering a systematic and scalable framework that reduces unbounded stylistic variation into a bounded set of functional categories.

B Average token-level confidence

We use the average token-level confidence to evaluate the model's confidence. We define the confidence score of register r as

$$\mathrm{Conf}(r) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{L_j}\sum_{i=1}^{L_j}\max_{w\in V} P_M\left(w \mid X_{r,<i}^{(j)}\right)\right),$$

where r is the register, V is the vocabulary, L_j is the length of the generated continuation for the rewritten opening sentence, and $X_{r,<i}^{(j)}$ refers to the concatenation of the rewritten opening sentence and the first $i-1$ tokens already generated in the continuation, M represents the target model.

Table 6: Register categories and abbreviations

Name	Abbr.	Name	Abbr.
Lyrical	LY	Encyclopedia article	en
Spoken	SP	Research article	ra
Interview	it	Description of a thing or person	dtp
Interactive discussion	ID	FAQ	fi
Narrative	NA	Legal terms & conditions	lt
News report	ne	**Opinion**	OP
Sports report	sr	Review	rv
Narrative blog	nb	Opinion blog	ob
How-to or instructions	HI	Denominational religious blog or sermon	rs
Recipe	re	Description with intent to sell	ds
Informational persuasion	IP	Informational description	IN
News & opinion blog or editorial	ed		

C Opening Template

Table 7 presents the representative opening templates T_r that we derived for each register. These templates were obtained by synthesizing a small subset of samples into the corresponding register r and then prompting a large language model to extract a generalized first-sentence structure. As shown in the table, each register exhibits distinct stylistic cues in its openings: for example, lyrical texts often begin with abstract imagery, interviews with a direct address from the interviewer, and storytelling narratives with a scene-setting phrase such as "Once upon a time." Such templates capture the prototypical entry points of different registers, which serve as useful anchors for aligning synthesized outputs with their intended discourse forms.

D Mathematical Notation

We summarize the key mathematical symbols used throughout this paper.

Symbol	Description
$\mathcal{D}_{\text{pro}}$	Proprietary dataset owned by the data rights holder.
$\mathcal{D}_{\text{held}}$	held-out dataset disjoint from $\mathcal{D}_{\text{pro}}$, used for validation.
M_t	Target model trained by the provider on laundered data.
M_a	Auxiliary LLM available to the data rights holder.
$\mathcal{R}$	Set of 23 established registers.
r^*	Register selected as most closely aligned with the laundering goal.
n, m, l	Sample sizes used in constructing templates, scoring, and inference, respectively.
K	Maximum number of iterations in the detailed prompt condition inference stage.
Standard-prompt$_r$	A canonical prompt that synthesizes text into register r.
T_r	Opening template extracted for register r.
$\hat{s}$	Synthetic data generated by M_a from an original sample s under prompt p.
$\tilde{s}$	Continuation produced by the target model M_t when prompted with $\hat{s}$.
$\text{Conf}(r)$	Average next-token confidence of M_t under register r.
$\mathcal{C}$	Candidate set of top-k registers with highest confidence scores.
Syn_r	Dataset synthesized into register r by M_a.
Perf_r	Unauthorized training data detection performance of Syn_r on M_t.
p	Current reverse-synthesis prompt refined during iterations.

Table 7: Registers and their corresponding opening templates

Register	Template (first-sentence / opening)
Lyrical	In the heart of [abstract domain], a tale unfolds, where [abstract concept], [abstract detail], [abstract entity], [abstract action].
Spoken style	So, let's talk about [TOPIC].
Interview	**Interviewer:** Thank you for joining us, [Person/Expert Title]. Can you tell us about [Subject/Topic]?
Interactive discussion	**[Participant 1]:** So, have you guys heard about [Topic/Subject]? I recently came across some interesting information about it.
Storytelling narrative	Once upon a time, in a [adjective] [type of place] called [place name], there lived a [adjective] [type of character] named [character name].
News report	**[Event/Topic]: [Description/Significance] [Location/Context]** – [Details about the subject, including noteworthy contributions, roles, or milestones].
Sports report	In a thrilling [event/display/action], [subject/actor] has [verb] [description/impact] in [field/area/genre].
Narrative blog post	In the context of [broad category or field], [subject or specific work] has made a significant impact, often leading to [general observation or effect].
Step-by-step guide (How-to)	**Step-by-Step Guide to Understanding [Subject]** — Step 1: [Initial focus or background]. Learn that [Subject Description].
Recipe	**Recipe for [General Concept]: [Specific Edition/Style]** — *Ingredients:* [Variable 1], [Variable 2], [Variable 3]...
Encyclopedia article	**[Subject]** is a [type/category] that [provides a description or function], [additional information if applicable].
Research article	This article explores the significance of [subject or topic], a [description or classification], characterized by [notable features or contributions].
Description of a thing or person	Introducing [Subject/Entity], a [descriptor] [type/category] [context/detail] renowned for its [property/characteristic].
FAQ	**What is [Subject]?** — [Subject] is a [general category or description] [specific type or detail] [additional information].
Legal terms & conditions	**Terms and Conditions Regarding [Subject/Theme]**.
Opinion	In my view, [Subject/Entity] represents [significance/impact/legacy] in [field/area], and its influence on [audience/community/context] cannot be overstated.
Review	[Subject] is a [descriptor] that [verb phrase] [contextual information].
Opinion blog (editorial)	When we think of [general category or field], [a notable example or subject] often comes to mind.
Denominational religious sermon	Beloved congregation, today we gather to reflect upon [individual/concept] that illuminates our lives and encourages us to contemplate our shared journey.
Description with intent to sell	Introducing [Subject]: a [descriptor] [product/service] designed for [use case]; discover how it [benefit/outcome] for [target user].
Informational persuasion	In the context of [domain or field], few [types/categories] resonate as profoundly within [subfields] as [specific work/name/entity].
Informational description	[Entity/Subject] is a [description] in the field of [broader category], specifically within [subcategory/locale].
News & opinion blog or editorial	When we think of [general category or field], [notable subject] often comes to mind — situating today's discussion of [topic] within [context].

E Evaluation Metrics

Following Carlini et al. (Carlini et al., 2022), we adopt three complementary metrics to evaluate membership inference attacks.

Area Under the ROC Curve (AUC). AUC measures the overall discriminative power of the attack, independent of any specific threshold. It reflects how well an unauthorized training data detection method can separate training data from unseen data on average. It may overstate effectiveness since it also includes high false-positive regions that are less relevant in practice.

Attack Success Rate (ASR). ASR measures the fraction of correctly identified training data under a single decision threshold that maximizes balanced accuracy across training data and unseen data. Unlike AUC, ASR reflects the practical effectiveness of an attack when deployed, as real-world unauthorized training data detentions typically operate at a single fixed threshold.

True Positive Rate at 5% False Positive Rate (TPR@5%). This metric evaluates the ability of a detection to identify training data while maintaining a strict false-positive constraint. Prior work highlights that low false-positive regimes are the most meaningful for privacy evaluation, since even a small number of incorrect training data decisions can undermine the credibility of the attack. TPR@5% therefore provides a high-precision view of attack success.

F Details of Implementation and Evaluation

We evaluate the effectiveness of our approach by examining whether the data reversed by **SDR** enhances the performance of existing unauthorized training data detection methods against LLMs trained on synthesized data. Specifically, we first sample 200 data points from the dataset that have not been seen by the *target* LLM. These 200 samples are synthesized into a new version using a prompt to simulate the data laundering process applied by the model provider. We then randomly select half of these synthesized samples as training data to fine-tune the LLM (training details are provided in Table 8), while the remaining half serves as non-training data. Subsequently, **SDR** is employed to infer a transformation prompt that restores the synthesis process and recovers the synthesized samples. Using this inferred prompt, we rewrite the original 200 data. Finally, we apply baseline unauthorized training data detections to both the original 200 samples and the 200 samples rewritten with the inferred prompt, comparing their attack performance to assess whether the inferred prompt improves the effectiveness of detections.

Lora_alpha	r	epochs	lr	gradient accumulation steps	weight_decay	logging_strategy
32	8	60	0.0004	25	0.1	"epoch"

Table 8: Hyperparameter settings for LoRA fine-tuning.

G Synthesized Prompts Used in Experiences

Table 9 lists the complete set of inside-register prompts corresponding to the 23 predefined registers in our framework. Each prompt is generated by GPT-5 using the instruction: "Generate a rewriting prompt that transfers the text into [REGISTER]." Here, [REGISTER] denotes one of the 23 registers. Table 10 lists the complete set of outside-register prompts; none align with any of the 23 predefined registers. These prompts are generated by GPT-5 using the instruction: "You are a prompt generator. Generate rewriting prompts that transform the original content into broadly different domains (e.g., Facebook post, academic article, children's story, legal document). Make the prompts as mutually distinct in domain as possible, and ensure each induces substantial changes to the original text rather than merely surface-level edits."

H The Specific Results for Each Prompt

Table 2 reports the average performance of each unauthorized training data detection method across models fine-tuned on data synthesized with different prompts. Tables 11 and 12 present the detailed average performance across five unauthorized training data detection methods for each inside-synthesized prompt. Tables 13 present the details for each outside-synthesized prompt. The "Original Prompt" column denotes the true synthesized prompt. The "Reversed Prompt" column denotes the best prompt recovered by **SDR** . The "Orig." column reports the average AUC across detection methods using the original data, whereas the "**SDR** " column reports the results using **SDR** -reversed data.

Table 9: Inside register prompts

ID	Synthesize Prompts
1	Rewrite the text in a lyrical style, ensuring the imagery is vivid, the rhythm flows naturally.
2	Rewrite the text in a spoken style, making it sound natural and conversational, and ensure the tone feels engaging and easy to follow for a live audience.
3	Rewrite the text in the form of an interview, ensuring the questions flow naturally and the answers provide clear, engaging explanations for the audience.
4	Rewrite the text as an interactive discussion between two or more participants, ensuring the conversation flows logically, with each speaker's tone and style clearly distinguishable.
5	Rewrite the text as a storytelling narrative. The story should flow naturally, use simple and engaging language, and be easy for all kinds of listeners to follow.
6	Rewrite the text in the style of a news report, ensuring the information is presented objectively and concisely.
7	Rewrite the text as a sports report, ensuring the action is described with dynamic, energetic language that conveys the pace, tension, and excitement of the event.
8	Rewrite the text as a narrative blog post, organized into clear sections with subheadings. Use a tone that is engaging and reflective, blending storytelling with explanation.
9	Rewrite the text as a step-by-step instructional guide. Break the content into numbered steps, with each step beginning with a clear imperative verb.
10	Rewrite the text as a recipe, introduce the information as sequential steps.
11	Rewrite the text to persuade the reader through factual information, making sure to include at least three specific data points or statistics to support the argument.
12	Rewrite the text as a sales description, and be sure to include a clear call-to-action at the end.
13	Rewrite the text in the style of an editorial, making sure to include a clear stance or opinion and a concluding paragraph that calls for action or reflection.
14	Rewrite the text as an informational description, ensuring the tone is neutral and objective, and include at least one definition or clarification to help the reader better understand the subject.
15	Rewrite the text in the style of an encyclopedia entry, maintaining a neutral, authoritative tone, and include at least one date, fact, or reference to give it the appearance of being sourced.
16	Rewrite the text as an academic research article, structured with sections such as Abstract, Introduction, Method, Results, and Conclusion, and include at least one in-text citation (invented if necessary) to simulate scholarly referencing.
17	Rewrite the text as a descriptive profile of a specific thing or person, using vivid details and attributes (appearance, characteristics, or context) and ending with a short summary sentence that highlights its significance.
18	Rewrite the text in the form of a Frequently Asked Questions (FAQ) section, making sure to include at least three question–answer pairs, with the questions phrased from the perspective of a curious reader.
19	Rewrite the text as legal terms and conditions, using formal legal language, and ensure at least one numbered clause is included for clarity.
20	Rewrite the text as a personal opinion piece, written in the first person, making sure to clearly express a stance and support it with at least one reason or example.
21	Rewrite the text as a review, giving it a clear positive or negative stance, and include at least one specific detail or example to justify the evaluation.
22	Rewrite the text as an opinion blog post, written in a conversational and persuasive tone, and include at least one personal anecdote or illustrative example to strengthen the argument.
23	Rewrite the text as a denominational religious sermon, using a reverent and exhortative tone, and include at least one scriptural quotation or moral teaching to guide the audience toward reflection or action.

Table 10: Outside register prompts

ID	Synthesize Prompts
1	Rewrite the following content as slide presentation bullet points. Focus on summarizing the key arguments and findings clearly and concisely. Use concise phrases that highlight core points.
2	Rewrite the following text in the style of a Facebook post. Sharing interesting information with followers. You may add light commentary, questions to the audience, or casual phrasing, but keep it natural and human-like. Avoid using emojis, hashtags, or overly dramatic expressions.
3	Adapt the text into a poetic form with vivid metaphors, rhythmic structure, and emotionally evocative language.
4	Convert the content into a tutorial-style explanation for beginners, using step-by-step instructions, simple analogies, and common misunderstandings.
5	Rewrite the text as a formal business email, ensuring clarity, professionalism, and a polite tone.
6	Rewrite the passage as a scientific abstract, including Background, Methods, Results, and Conclusions. Invent at least two numerical values (percentages, sample sizes, or statistical outcomes) to support claims.
7	Rewrite the text as a product description for an e-commerce website, highlighting key features, benefits, and use cases in a persuasive manner.
8	Rewrite the text as a blog post, incorporating vivid descriptions of locations, cultural insights, and personal experiences to engage readers.
9	Rewrite the text as a classroom lecture transcript, with explanations, rhetorical questions, and occasional student interaction.
10	Rewrite the text with stronger transitions between sentences and paragraphs, ensuring smoother reading without adding new information.

H.1 Analysis of non-goal synthesis process

We analyze a special case where the synthesis prompt does not provide an explicit directive. For example, consider the outside-synthesized prompt: "Rewrite the text with stronger transitions between sentences and paragraphs, ensuring smoother reading without adding new information." Although this instruction lacks a clear goal keyword, **`SDR`** infers a broader editorial-style prompt: "Rewrite the text in the style of an editorial, focusing on enhancing the narrative through emotional engagement, historical significance, and the subject's impact, while highlighting community involvement and contemporary relevance." Despite the absence of an explicit goal, this inferred prompt enhances the performance of unartificialized training data detection, particularly improving the AUC of Loss (Yeom et al., 2018) from 0.538 to 0.669.

I Limitation

A key limitation of our current approach is that the register taxonomy it relies on is too coarse-grained to locate goals accurately. Although the existing taxonomy of 23 sub-registers offers broad coverage of textual styles, it was not initially designed for classifying laundering goal. Consequently, there are cases where none of the 23 registers can adequately capture the intent of a synthesized prompt, leading to reduced accuracy in goal identification and, in turn, lower quality in restored prompts. Overcoming this limitation calls for future research on developing more fine-grained taxonomies tailored to synthesized data, thereby enabling more accurate and robust prompt reversal in practical scenarios.

J AI usage clarification

Large Language Models improved the manuscript's grammar and readability; all research design, analysis, and interpretation were conducted by the authors.

Table 11: Inside register prompts and corresponding reversed prompts with Orig and **SDR** results.

ID	Original Prompt	Reversed Prompt (SDR)	Orig	SDR
1	Rewrite the text in a lyrical style, ensuring the imagery is vivid, the rhythm flows naturally.	Rewrite the text in a lyrical style, enhancing the poetic rhythm and imagery while capturing the essence and emotional depth of the original content.	0.540	0.692
2	Rewrite the text in a spoken style, making it sound natural and conversational, and ensure the tone feels engaging and easy to follow for a live audience.	Rewrite the text to sound natural and conversational, using everyday language and personal anecdotes to create an engaging and friendly atmosphere for the listener.	0.702	0.894
3	Rewrite the text in the form of an interview, ensuring the questions flow naturally and the answers provide clear, engaging explanations for the audience.	Rewrite the text in the form of an interview, ensuring a clear and engaging dialogue that accurately conveys the information while maintaining a conversational tone and eliciting detailed responses.	0.713	0.823
4	Rewrite the text as an interactive discussion between two or more participants, ensuring the conversation flows logically, with each speaker's tone and style clearly distinguishable.	Rewrite the text as an interactive discussion between two or more participants, ensuring a natural flow of dialogue that incorporates factual information and engages with the topic through building on each other's comments.	0.650	0.770
5	Rewrite the text as a storytelling narrative. The story should flow naturally, use simple and engaging language, and be easy for all kinds of listeners to follow.	Rewrite the text in the form of a Frequently Asked Questions section, transforming the information into clear and concise questions and answers that emphasize key details and engage the reader effectively.	0.646	0.735
6	Rewrite the text in the style of a news report, ensuring the information is presented objectively and concisely.	Rewrite the text into a Frequently Asked Questions section, organizing the information into clear and concise questions and answers while highlighting key details and maintaining clarity and readability.	0.793	0.904
7	Rewrite the text as a sports report, ensuring the action is described with dynamic, energetic language that conveys the pace, tension, and excitement of the event.	Rewrite the text in the style of an engaging editorial, enhancing the narrative through vivid language, emotional depth, and a focus on the significance of the subject matter.	0.655	0.713
8	Rewrite the text as a narrative blog post, organized into clear sections with subheadings. Use a tone that is engaging and reflective, blending storytelling with explanation.	Rewrite the text in the form of a Frequently Asked Questions section, focusing on clearly structured questions and answers that highlight key aspects, contributions, and significance of the subject matter in a conversational tone.	0.667	0.748
9	Rewrite the text as a step-by-step instructional guide. Break the content into numbered steps, with each step beginning with a clear imperative verb.	Rewrite the text as a step-by-step instructional guide, breaking down the information into clear, organized steps that highlight key concepts, details, and relevant insights for enhanced understanding.	0.778	0.841
10	Rewrite the text as a recipe, introduce the information as sequential steps.	Rewrite the text to persuasively present factual information, emphasizing key aspects and structuring the content clearly to enhance engagement and clarity.	0.700	0.724
11	Rewrite the text to persuade the reader through factual information, making sure to include at least three specific data points or statistics to support the argument.	Rewrite the text in the style of an encyclopedia entry, focusing on enhancing structural clarity, coherence, and technical detail by organizing the information into distinct sections and emphasizing historical significance and key contributions.	0.633	0.633
12	Rewrite the text as a sales description, and be sure to include a clear call-to-action at the end.	Rewrite the text as a Frequently Asked Questions section, transforming the original content into a clear and engaging question-and-answer format that effectively highlights key elements, significance, and context for the reader.	0.622	0.725

Table 12: Inside register prompts and corresponding reversed prompts with Orig and **SDR** results (continues).

ID	Original Prompt	Reversed Prompt (SDR)	Orig	SDR
13	Rewrite the text in the style of an editorial, making sure to include a clear stance or opinion and a concluding paragraph that calls for action or reflection.	Rewrite the text as a personal opinion piece, emphasizing reflective commentary and personal insights while exploring the broader societal implications and significance of the subject matter.	0.594	0.657
14	Rewrite the text as an informational description, ensuring the tone is neutral and objective, and include at least one definition or clarification to help the reader better understand the subject.	Rewrite the text as a step-by-step instructional guide, organizing the content into clear, numbered sections that effectively communicate essential information about the subject.	0.594	0.795
15	Rewrite the text in the style of an encyclopedia entry, maintaining a neutral, authoritative tone, and include at least one date, fact, or reference to give it the appearance of being sourced.	Rewrite the text as an informational description, focusing on presenting a clear, structured overview of the subject's key facts, achievements, and background while maintaining concise and objective language.	0.792	0.923
16	Rewrite the text as an academic research article, structured with sections such as Abstract, Introduction, Method, Results, and Conclusion, and include at least one in-text citation (invented if necessary) to simulate scholarly referencing.	Rewrite the text in the form of an interview, transforming the original content into a conversational dialogue that incorporates engaging questions and responses while maintaining clarity and coherence.	0.525	0.546
17	Rewrite the text as a descriptive profile of a specific thing or person, using vivid details and attributes (appearance, characteristics, or context) and ending with a short summary sentence that highlights its significance.	Rewrite the text as a sales description, transforming it into an engaging narrative that highlights the subject's achievements, legacy, and emotional impact to captivate and appeal to potential audiences.	0.661	0.765
18	Rewrite the text in the form of a Frequently Asked Questions (FAQ) section, making sure to include at least three question–answer pairs, with the questions phrased from the perspective of a curious reader.	Rewrite the text in the form of an interview, transforming factual information into a conversational question-and-answer format that captures personal insights, key themes, and details from the original content.	0.640	0.792
19	Rewrite the text as legal terms and conditions, using formal legal language, and ensure at least one numbered clause is included for clarity.	Rewrite the text into a Frequently Asked Questions section by converting the content into clear questions and answers, ensuring clarity, conciseness, accuracy, and structured organization of information.	0.731	0.833
20	Rewrite the text as a personal opinion piece, written in the first person, making sure to clearly express a stance and support it with at least one reason or example.	Rewrite the text as a review, focusing on summarizing key aspects and implications while maintaining an engaging narrative style that connects with the reader.	0.575	0.613
21	Rewrite the text as a review, giving it a clear positive or negative stance, and include at least one specific detail or example to justify the evaluation.	Rewrite the text as a review, emphasizing the subject's significance, key achievements, and connections to broader themes or contexts, while maintaining a consistent tone and providing a balanced evaluation.	0.593	0.698
22	Rewrite the text as an opinion blog post, written in a conversational and persuasive tone, and include at least one personal anecdote or illustrative example to strengthen the argument.	Rewrite the text as a conversational interview, focusing on transforming factual content into dialogue by incorporating questions, responses, and personal insights while maintaining the original essence.	0.577	0.630
23	Rewrite the text as a denominational religious sermon, using a reverent and exhortative tone, and include at least one scriptural quotation or moral teaching to guide the audience toward reflection or action.	Rewrite the text as a denominational religious sermon, transforming the narrative into an inspirational message that emphasizes spiritual themes, fosters community, and resonates with the congregation's values.	0.541	0.627

Table 13: Outside register prompts and corresponding reversed prompts with Orig and **SDR** results.

ID	Original Prompt	Reversed Prompt (SDR)	Orig	SDR
1	Rewrite the following content as slide presentation bullet points. Focus on summarizing the key arguments and findings clearly and concisely. Use concise phrases that highlight core points.	Rewrite the text as a step-by-step instructional guide, organizing the information into clear sections and ensuring each step provides concise, relevant details on the specified topic.	0.730	0.799
2	Rewrite the following text in the style of a Facebook post. Sharing interesting information with followers. You may add light commentary, questions to the audience, or casual phrasing, but keep it natural and human-like. Avoid using emojis, hashtags, or overly dramatic expressions.	Rewrite the text as a Frequently Asked Questions section, transforming the information into an engaging question-and-answer format that encourages reader interaction and maintains a conversational tone.	0.644	0.730
3	Adapt the text into a poetic form with vivid metaphors, rhythmic structure, and emotionally evocative language.	Rewrite the text in a lyrical style that transforms factual content into an evocative narrative, using vivid imagery, poetic devices, and rhythmic flow to highlight emotional resonance and thematic cohesion.	0.538	0.674
4	Convert the content into a tutorial-style explanation for beginners, using step-by-step instructions, simple analogies, and common misunderstandings.	Rewrite the text as a step-by-step instructional guide, ensuring clear and concise steps that effectively outline key aspects and concepts while maintaining an engaging tone and logical flow throughout.	0.616	0.679
5	Rewrite the text as a formal business email, ensuring clarity, professionalism, and a polite tone.	Rewrite the text in the style of an encyclopedia entry, emphasizing clear and concise organization, formal language, and distinct sections that present factual information and key points effectively.	0.795	0.884
6	Rewrite the passage as a scientific abstract, including Background, Methods, Results, and Conclusions. Invent at least two numerical values (percentages, sample sizes, or statistical outcomes) to support claims.	Rewrite the text as a Frequently Asked Questions section, transforming the original content into clear, concise questions and answers that emphasize key themes, significant information, and factual accuracy.	0.600	0.678
7	Rewrite the text as a product description for an e-commerce website, highlighting key features, benefits, and use cases in a persuasive manner.	Rewrite the text as a dialogue in an interview format, emphasizing key details and insights while maintaining clarity and engagement through a question-and-answer structure.	0.603	0.673
8	Rewrite the text as a blog post, incorporating vivid descriptions of locations, cultural insights, and personal experiences to engage readers.	Rewrite the text as a sales description that emphasizes unique aspects and engaging narratives, highlighting significance and emotional appeal to captivate the audience.	0.626	0.728
9	Rewrite the text as a classroom lecture transcript, with explanations, rhetorical questions, and occasional student interaction.	Rewrite the text in the form of an interview, transforming the information into a natural dialogue that incorporates questions and answers while preserving the original content's key details and themes.	0.667	0.764
10	Rewrite the text with stronger transitions between sentences and paragraphs, ensuring smoother reading without adding new information.	Rewrite the text in the style of an editorial, focusing on enhancing the narrative through emotional engagement, historical significance, and the subject's impact, while highlighting community involvement and contemporary relevance.	0.570	0.646