GSPS: GLOBAL STYLE-GUIDED PROTECTION SYNTHESIS FOR OWNERSHIP VERIFICATION AND LEAKAGE TRACING OF UNLABELED DATASETS IN SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Curated unlabeled datasets are valuable assets for semi-supervised learning (SSL), yet remain vulnerable to unauthorized copying, redistribution, and reuse. Protecting such datasets is challenging because conventional watermarking methods often rely on stable label supervision, while SSL operates largely on unlabeled data and employs strong augmentations that can disrupt watermark signals. Moreover, existing methods mainly support binary ownership verification and cannot trace leaked copies to specific recipients under black-box auditing. To address these challenges, we propose **GSPS**, a **Global Style-guided Protection Synthesis** framework for ownership verification and recipient-specific leakage tracing of unlabeled text datasets. GSPS first identifies reliable watermark carriers without human annotations using EFS-χ^2-based selection for text classification and multi-model cooperation for LLM-based reasoning and question answering. It then introduces *Quad-Dimensional Adaptive Style Selection* (QASS) to identify robust global watermark styles and integrates semantically separated recipient-specific fingerprints into the selected style to generate distinguishable dataset copies. During black-box auditing, GSPS performs statistical ownership verification and recipient-specific leakage attribution. Extensive experiments on Chinese and English datasets, multiple SSL algorithms, and both text classification and LLM-based reasoning and question-answering tasks demonstrate that GSPS achieves up to 90% watermark success rate and over 95% trace accuracy at only a 1% watermarking rate, with negligible utility degradation and strong robustness against state-of-the-art defenses.

1 INTRODUCTION

Semi-supervised learning (SSL) has emerged as an effective paradigm for reducing the reliance of language models on large-scale human annotations. By jointly leveraging a limited amount of labeled data and abundant unlabeled data, SSL has demonstrated strong performance across a wide range of tasks, including natural language understanding, text classification, reasoning, and question answering Liang (2005); Xie et al. (2020); Zou & Caragea (2023); Luo et al. (2025). Although such data avoid annotation costs, their collection, filtering, deduplication, and quality control remain expensive. Consequently, curated unlabeled datasets have become valuable assets whose unauthorized reuse can incur substantial economic and intellectual-property risks. Recent advances in large-scale language models further highlight the critical role of data quality and scale in shaping model capabilities Zha et al. (2025); Mazumder et al. (2023); Ghorbani & Zou (2019); Yoon et al. (2020), making carefully curated datasets indispensable resources for model training and adaptation Kaplan et al. (2020); Brown et al. (2020); Raffel et al. (2020); Bommasani et al. (2021).

Despite their growing value, the ownership of unlabeled datasets remains largely unprotected. Digital datasets can be easily copied, redistributed, and reused without authorization, enabling third parties to train competitive models at minimal cost. While prior IP protection studies have developed watermarking and fingerprinting techniques for models and datasets Zhang et al. (2021b); Zheng et al. (2022); Lukas et al. (2019); Jia et al. (2021); Li et al. (2022b; 2023), binary owner-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ship verification alone is insufficient in practical distribution scenarios. In commercial licensing or collaborative training, the same proprietary dataset is often distributed to multiple authorized recipients. Once a leaked copy is used to train an unauthorized model, the dataset owner must answer two distinct questions: whether the suspicious model was trained on the protected data, and, if so, which recipient leaked it. A shared watermark only addresses the former and cannot provide recipient-level attribution. Leakage tracing therefore requires recipient-specific signals that remain distinguishable after training and observable through black-box outputs. This necessitates protection mechanisms that jointly support ownership verification and recipient-specific leakage tracing under realistic black-box auditing Li et al. (2023); Wegerhoff et al. (2024).

Extending existing dataset watermarking techniques to unlabeled SSL data, however, introduces several fundamental challenges. **First, label-free watermark embedding is non-trivial.** Traditional dataset watermarking relies on ground-truth labels to bind triggers to a target class Sablayrolles et al. (2020); Li et al. (2022a); Tang et al. (2023); Li et al. (2023); Guo et al. (2024), but unlabeled SSL data has no true labels to establish this association; moreover, as model predictions shift during self-training, pseudo-labels may fluctuate across classes or fall below confidence thresholds, causing watermark carriers to receive inconsistent supervision or to be excluded from training altogether Sohn et al. (2020); Zhang et al. (2021a); Wang et al. (2022). Consequently, conventional label-dependent methods cannot be directly applied without additional annotations. **Second, watermark signals must survive both SSL augmentations and adaptive defenses.** Modern SSL methods commonly enforce prediction consistency under strong data augmentations, whereas token- or phrase-level triggers can be disrupted by transformations such as back-translation and word-order perturbation, thereby weakening the learned trigger-response association Chen et al. (2021); Qi et al. (2021c); Yan et al. (2023). Global style-based triggers provide a promising alternative because they distribute the watermark signal across broader linguistic patterns rather than isolated tokens. However, existing style-based methods are not explicitly optimized for SSL augmentations and remain vulnerable to targeted defenses Qi et al. (2021b); Pan et al. (2022); Qi et al. (2021a); Yang et al. (2021a). **Third, recipient-level tracing requires fingerprints with stronger separability than conventional ownership watermarks.** Multiple recipient-specific signals must remain distinguishable after SSL training and potential defenses, while still being detectable solely through black-box outputs Li et al. (2023); Wegerhoff et al. (2024).

To address these challenges, we propose ***GSPS***, a ***Global Style-guided Protection Synthesis*** framework for joint ownership verification and recipient-level leakage tracing of unlabeled text datasets in SSL. GSPS addresses the three challenges through a unified protection pipeline. First, it selects reliable watermark carriers directly from unlabeled data, reducing the impact of pseudo-label instability during SSL training. Second, it combines a shared global style for ownership verification with recipient-specific fingerprints for leakage attribution, balancing common ownership evidence with cross-recipient distinguishability. Third, it enables black-box auditing through statistical ownership testing and recipient-level attribution under both probability-accessible and label-only outputs. This design enables unified ownership protection and leakage tracing across text classification, reasoning, and question-answering tasks.

We summarize our contributions as follows: **1)** To the best of our knowledge, GSPS is the first unified framework to jointly support ownership verification and recipient-level leakage tracing for unlabeled text datasets under semi-supervised learning, enabling practical auditing under both probability-accessible and label-only black-box settings. **2)** We develop a label-free watermark carrier selection strategy that identifies target-aligned unlabeled samples without ground-truth annotations, thereby mitigating pseudo-label instability during semi-supervised training. **3)** We introduce Quad-Dimensional Adaptive Style Selection (QASS) to discover robust global styles that serve as shared ownership signals, and further incorporate semantically separated recipient-specific fingerprints to enable reliable leakage attribution. Extensive experiments across 13 Chinese and English benchmarks demonstrate watermark success rates of up to 91.9% at only a 1% watermarking rate, while achieving 97.5–100.0% tracing accuracy under probability-accessible auditing and 92.8–98.2% under label-only auditing, with negligible utility degradation and strong robustness against representative defenses.

2 Related Work

2.1 Semi-Supervised Learning for NLP

Semi-supervised learning (SSL) mitigates annotation scarcity by coupling a small labeled dataset with abundant unlabeled text Liang (2005); Xie et al. (2020); Zou & Caragea (2023). Modern semi-supervised fine-tuning protocols integrate empirical risk minimization on labeled samples with structural regularization over unlabeled distributions via dynamic pseudo-labeling and consistency optimization Zhang et al. (2021a); Wang et al. (2022); Zou & Caragea (2023); Luo et al. (2025).

For text classification, mainstream SSL frameworks primarily rely on data augmentations and iterative pseudo-label updates. Consistency-based paradigms, such as UDA and FixMatch, enforce prediction consistency across weakly and strongly augmented views of identical inputs Xie et al. (2020); Sohn et al. (2020). To mitigate confirmation bias and error accumulation, subsequent approaches introduce adaptive confidence estimation: FlexMatch and FreeMatch dynamically adjust class-specific selection thresholds, while collaborative strategies like JointMatch refine multi-perspective pseudo-labeling Zhang et al. (2021a); Wang et al. (2022); Zou & Caragea (2023).

For LLM reasoning and question-answering (QA) tasks, semi-supervised fine-tuning protocols exploit the contextual reasoning capabilities of foundation models Brown et al. (2020); Touvron et al. (2023); Bommasani et al. (2021), integrating unannotated instances directly into downstream fine-tuning loops Gururangan et al. (2020); Raffel et al. (2020); Luo et al. (2025). In particular, frameworks such as SEMIEVOL utilize iterative self-refinement and pseudo-reasoning paths under strict annotation budgets Wang et al. (2023); Luo et al. (2025). However, continuous pseudo-label evolution and aggressive data perturbations across these SSL workflows inherently dilute static signals, presenting fundamental challenges to dataset ownership protection.

2.2 Backdoor Attacks in NLP

Textual backdoor attacks inject covert triggers into training data to force a target prediction on triggered inputs while maintaining clean accuracy Chen et al. (2021); Qi et al. (2021c). Early paradigms relied on static character-, word-, or sentence-level insertions Chen et al. (2021); Zheng et al. (2025). To improve imperceptibility, subsequent works developed stealthy triggers: StyleBkd applied neural style transfer to embed syntactic attributes, whereas Hidden Killer and LWS leveraged syntactic parse trees or dynamic substitution spaces as invisible signals Qi et al. (2021b); Pan et al. (2022); Qi et al. (2021c;d).

Recent research has further explored context-aware and clean-label mechanisms to enhance fluency and stealthiness Yan et al. (2023); Gan et al. (2022). For instance, BITE introduces LM-guided word perturbations, CL-Attack constructs cross-lingual structures, and Triggerless Backdoor Attack optimizes poison instances in representation space without label flipping Yan et al. (2023); Zheng et al. (2025); Gan et al. (2022); Zhao et al. (2023; 2024). However, these architectures are predominantly designed for supervised learning. Under semi-supervised dynamics, aggressive data augmentations disrupt localized patterns, while iterative pseudo-label updates continuously overwrite static poison signals Xie et al. (2020); Sohn et al. (2020); Zhang et al. (2021a).

2.3 Dataset Copyright Protection

Dataset ownership protection embeds persistent watermarks into distributed data, enabling owners to audit unauthorized usage via black-box hypothesis testing on suspect model outputs Sablayrolles et al. (2020); Li et al. (2023); Tang et al. (2023); Nie & Yuan (2025); Zheng et al. (2022). A prevalent paradigm repurposes backdoor mechanisms to inject unique trigger-label shortcuts Li et al. (2023); Tang et al. (2023); Jia et al. (2021). Recent extensions investigate zero-leakage verification without disclosing triggers (*e.g.*, ZeroMark Guo et al. (2024)), compression resistance Nie & Yuan (2025), pre-trained backbone tracking Xie et al. (2025), and multi-recipient leaker identification (*e.g.*, DataDetective Wegerhoff et al. (2024)).

Despite recent empirical progress, existing dataset protection schemes remain difficult to apply directly to semi-supervised settings. White-box approaches, such as EaaW Zhang et al. (2021b) and EmbTracker Li et al. (2022b), require access to model gradients, parameters, or internal representa-

tions, making them unsuitable for realistic auditing scenarios in which only black-box API access is available. Meanwhile, conventional label-dependent backdoor watermarking and supervised tracing methods rely on predefined trigger–label associations, which become unreliable in SSL because unlabeled samples are governed by dynamically evolving pseudo-labels. Such pseudo-label fluctuations can introduce inconsistent supervision or exclude watermark carriers from training, substantially weakening the learned watermark signal. Recent paraphrase- and output-level watermarking methods (*e.g.*, Shetty et al. (2026); Dabiriaghdam & Wang (2025)) improve the protection of generated content or model APIs, but are not designed to construct reliable watermark signals from unlabeled SSL data or to support scalable recipient-level attribution. In contrast, GSPS is specifically designed for unlabeled SSL settings and requires only black-box model outputs for auditing. By combining label-free watermark construction with shared ownership signals and recipient-specific fingerprints, GSPS jointly enables statistically grounded ownership verification and multi-recipient leakage attribution while mitigating watermark degradation caused by dynamic pseudo-labeling.

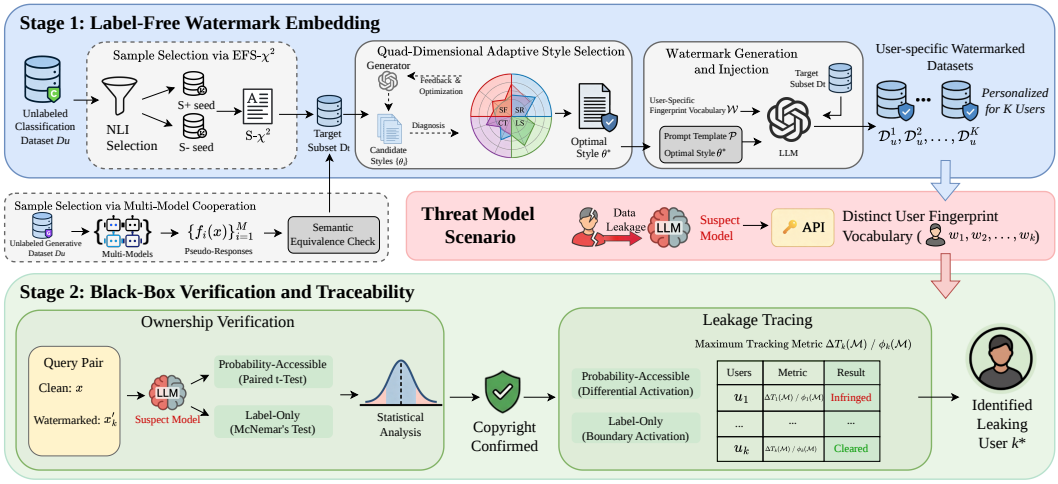


Figure 1: Overview of the proposed GSPS framework.

3 Methodology

3.1 Problem Formulation

We formalize the threat model and introduce the foundational notation for dataset watermarking and leakage tracing in semi-supervised learning settings.

3.1.1 Threat Model

Data Owner's Goal and Capability The dataset owner seeks to detect unauthorized use of proprietary unlabeled data and, when misuse is identified, attribute the leakage to a specific authorized recipient without relying on ground-truth annotations. To this end, the owner embeds a shared global style-based watermark together with recipient-specific identifiers into a small fraction of the unlabeled dataset, producing customized protected copies for different recipients while preserving data quality, watermark stealthiness, and downstream task utility.

For ownership verification and leakage attribution, we assume that the owner has only black-box query access to a suspicious downstream model, under either probability-accessible or label-only settings. The auditing process relies exclusively on standard input–output interactions and requires no access to model parameters, gradients, internal representations, network architectures, training data, or training configurations. This setting reflects realistic deployment scenarios in which suspicious models are exposed only through prediction APIs.

Adversaries' Goal and Capability We consider two complementary adversarial roles. First, an authorized but potentially malicious recipient may violate the data-use agreement by leaking, re-

selling, or redistributing its customized dataset copy to unauthorized parties. Second, an unauthorized downstream adversary may acquire the leaked copy and combine it with a limited amount of labeled data to train a target model using semi-supervised learning algorithms Xie et al. (2020); Luo et al. (2025). The adversary's objective is to exploit the protected data for downstream training while avoiding subsequent ownership verification and recipient-level attribution.

We further consider adversaries that may actively attempt to weaken or remove the embedded protection signals while maintaining downstream model utility. Such countermeasures may be applied before, during, or after training, including data-level sanitization and transformation (e.g., ONION and back-translation), style-neutralizing rewriting, anomaly-based filtering of suspicious samples, multi-recipient collusion, and model-level purification or adaptation (e.g., Fine-Pruning and supervised fine-tuning). These capabilities provide a strong threat model for evaluating whether the embedded ownership and recipient-specific signals remain detectable after realistic data manipulation, semi-supervised training, and post-training defenses Liu et al. (2018); Qi et al. (2021a); Yang et al. (2021a); Azizi et al. (2021).

3.1.2 PRELIMINARIES

Let $\mathcal{X}$ denote the input text space and $\mathcal{Y}$ be the task label space. Let $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ denote the total training dataset comprising a small labeled set $\mathcal{D}_l$ and a large-scale unlabeled set $\mathcal{D}_u = \{x_i\}_{i=1}^{|\mathcal{D}_u|} \subset \mathcal{X}$, where $\mathcal{D}_l \cap \mathcal{D}_u = \emptyset$. Let $\mathcal{U} = \{u_1, \ldots, u_K\}$ denote the set of K authorized recipients. We define the target watermarking rate as $\gamma = \frac{|\mathcal{D}_t|}{|\mathcal{D}|} = \frac{|\mathcal{D}_t|}{|\mathcal{D}_l|+|\mathcal{D}_u|} \ll 1$, where the target subset $\mathcal{D}_t \subset \mathcal{D}_u$ is curated exclusively from unannotated instances. For each recipient u_k, target instances $x \in \mathcal{D}_t$ are transformed into watermarked counterparts $x' = \mathcal{G}(x \mid \mathcal{P}(\theta^*, \mathcal{W}_k))$ using a text generator $\mathcal{G}$ guided by the optimal global style θ^* and the recipient's unique fingerprint token set $\mathcal{W}_k \subset \mathcal{W}_{\text{valid}}$. To resist cross-copy differential comparisons, we additionally apply label-invariant synonym cloaking $\mathcal{C}(\cdot)$ to an independent benign subset $\mathcal{D}_c^{(k)} \subset (\mathcal{D}_u \setminus \mathcal{D}_t)$, defining the customized dataset as $\mathcal{D}_w^{(k)} = \{x' \mid x \in \mathcal{D}_t\} \cup \{\mathcal{C}(x) \mid x \in \mathcal{D}_c^{(k)}\} \cup (\mathcal{D}_u \setminus (\mathcal{D}_t \cup \mathcal{D}_c^{(k)}))$. A suspect downstream model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ trained on $\mathcal{D}_l \cup \mathcal{D}_w^{(k)}$ implicitly captures the embedded watermark representations. Forensic investigation subsequently executes ownership verification via a decision rule $\mathcal{V}(\mathcal{M}) \in \{0, 1\}$ and leakage tracing via an attribution mapping $\mathcal{T}(\mathcal{M}) \in \{1, \ldots, K\} \cup \{\perp\}$, where $\perp$ denotes an inconclusive outcome.

3.2 THE GSPS FRAMEWORK

As illustrated in Fig. 1, GSPS consists of two stages: *label-free watermark embedding* and *black-box verification and traceability*. In Stage 1, GSPS first selects reliable watermark carriers from unlabeled data using EFS-χ^2 for text classification or multi-model cooperation for reasoning and question-answering tasks. It then identifies a robust global style through Quad-Dimensional Adaptive Style Selection (QASS) and incorporates recipient-specific fingerprints during watermark generation, producing personalized protected datasets for different authorized recipients. In Stage 2, the owner audits a suspicious downstream model through black-box queries under either probability-accessible or label-only API settings. Statistical hypothesis testing is first employed to verify whether the protected dataset has been used, after which recipient-specific fingerprint responses are compared to attribute the leakage to the most likely recipient.

3.2.1 HIGH-EFFICIENCY SAMPLE SELECTION

In semi-supervised learning, effective watermark injection requires selecting unlabeled samples that are reliably associated with the target class y_t. Randomly selecting unlabeled samples may include samples from different classes, making it difficult for the model to learn a clear association between the watermark and y_t. Moreover, incorrect or low-confidence pseudo-labels may cause watermarked samples to be assigned to the wrong class or removed during confidence-based filtering, further weakening the watermark signal. Therefore, we develop label-free sample selection strategies to identify high-confidence, target-aligned samples before watermark injection.

Classification Tasks. Pseudo-labeling an entire unannotated corpus via large language models (LLMs) is computationally prohibitive. More critically, LLMs readily overfit to isolated surface

words rather than full sentence context: models blindly assign high-confidence target labels upon encountering specific keywords while disregarding the overall semantics, which severely degrades pseudo-label reliability. To address these limitations, we propose EFS-χ^2, a seed-to-corpus bootstrapping framework that expands a compact set of extreme-confidence seed instances into globally invariant lexical statistics, enabling efficient and high-purity target carrier selection.

1) Target Seed Identification via Entropy Filtering: To eliminate prior token-frequency and format biases of zero-shot NLI models Liu et al. (2019) without labeled data, raw logits $l(x, h_y)$ are first calibrated against the empty premise prior $l(\emptyset, h_y)$ Zhao et al. (2021):

$$P(y \mid x) = \mathrm{Softmax}\big(l(x, h_y) - l(\emptyset, h_y)\big). \tag{1}$$

By evaluating calibrated predictive entropy across $\mathcal{D}_u$, a small set of extreme-confidence instances within the lowest 5% entropy partition is harvested as reference seeds: target-class seeds $\mathcal{S}^+$ ($y = y_t$) and non-target seeds $\mathcal{S}^-$ ($y \neq y_t$). These seeds serve as a high-confidence statistical baseline without manual annotations for computing global lexical Chi-Square statistics in the next step.

2) High-efficiency Sample Selection via Signed χ^2 Scoring: Although the identified seed sets exhibit high purity, their limited quantity cannot directly fulfill carrier requirements. To expand this high-confidence signal to the entire corpus, we leverage $\mathcal{S}^+$ and $\mathcal{S}^-$ to construct a global lexical discriminator S-$\chi^2(w)$ for each token w. All candidate instances $x \in \mathcal{D}_u$ are then evaluated via length-normalized polarized aggregation:

$$\mathrm{Score}(x) = \frac{1}{\sqrt{|x|}} \sum_{w \in x} \Phi\big(\text{S-}\chi^2(w)\big), \tag{2}$$

where $\Phi(\cdot)$ applies an asymmetric penalty to conflicting non-target tokens. Ranking $\mathcal{D}_u$ by $\mathrm{Score}(x)$ extracts the top-N_t ($N_t = \gamma|\mathcal{D}|$) highest-scoring target instances to form the final carrier set $\mathcal{D}_t$, with full pipeline details provided in Appendix A.1.

LLM-Based Reasoning and Question Answering Tasks. For unannotated LLM reasoning and question answering tasks, single-model pseudo-labeling often generates flawed yet overconfident rationales, mistakenly assigning non-target instances to target class y_t. Injecting watermarks into such misclassified samples corrupts subsequent self-training through confirmation bias. To eliminate these spurious carriers, we deploy a multi-model dual-filtering mechanism that retains an unannotated query only when multiple distinct models reach high-confidence consensus on target class y_t.

1) Candidate Identification via Multi-Model Agreement: To filter out ambiguous queries, we evaluate cross-model consensus across an ensemble of M diverse open-source LLMs $\{f_m\}_{m=1}^{M}$ (setup detailed in Appendix A.1) via $\mathrm{MA}(x) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}\big(f_m(x) = y_t\big)$. Only instances achieving unanimous consensus ($\mathrm{MA}(x) = 1.0$) are retained as candidate samples $\mathcal{D}_{\mathrm{cand}}$, eliminating subjective or model-specific labeling hallucinations.

2) Sample Selection via Predictive Confidence: Since consensus can occasionally occur on borderline queries, we further score instances in $\mathcal{D}_{\mathrm{cand}}$ using ensemble mean predictive confidence $\mathrm{MC}(x) = \frac{1}{M}\sum_{m=1}^{M} P_{f_m}(y_t \mid x)$, where $P_{f_m}(y_t \mid x)$ denotes the normalized target-token probability emitted by model f_m. Sorting $\mathcal{D}_{\mathrm{cand}}$ by descending $\mathrm{MC}(x)$ extracts the top-N_t ($N_t = \gamma|\mathcal{D}|$) queries with the highest semantic certainty to form the final target sample set $\mathcal{D}_t$.

3.2.2 Quad-Dimensional Adaptive Style Selection

Conventional style backdoors like StyleBkd Qi et al. (2021b) rely on static presets (e.g., Biblical or Shakespearean prose) that underperform in SSL by distorting semantics, introducing syntactic artifacts, and degrading under back-translation, which prevents models from learning cohesive watermark associations. To address this, QASS dynamically evaluates candidate style directives θ across four complementary dimensions via a generative LLM $\mathcal{G}$ and an embedding encoder $\mathbf{e}(\cdot)$ to identify the optimal directive θ^*: **(1) Semantic Fidelity** (SF)**:** Measures whether style transfer preserves the original meaning, based on textual entailment consistency and embedding similarity between the original and styled texts. **(2) Linguistic Stealthiness** (LS)**:** Measures whether the styled text remains linguistically natural, based on relative perplexity deviation and formal linguistic acceptability. **(3) Style Robustness** (SR)**:** Measures whether the style-induced representation

shift remains stable after common SSL text augmentations. **(4) Cluster Tightness** (CT)**:** Measures whether applying the same style directive produces consistent representation shifts across different samples.

The four dimensions are normalized into $[0, 1]$ and averaged into a composite fitness score $\mathrm{CS}(\theta) = \frac{1}{4}(\mathrm{SF} + \mathrm{LS} + \mathrm{SR} + \mathrm{CT})$. We use equal weighting as a model-agnostic default rather than claiming that the four dimensions contribute equally. Serving as both text generator and evolutionary agent, $\mathcal{G}$ receives reflective numerical feedback to iteratively mutate candidate style prompts toward the optimal directive $\theta^* = \arg\max_\theta \mathrm{CS}(\theta)$. As shown in Figures 2 and 3, this search converges within 20 iterations, and higher fitness is associated with stronger downstream WSR. Formulations and prompt templates are detailed in Appendices A.2 and D.1.

3.2.3 Watermark Generation and Injection

To enable ownership verification and leak tracing across multiple recipients, GSPS introduces a two-step pipeline: *Dual-Filter Fingerprint Construction* and *Fingerprint Injection*. Rather than inserting isolated tokens, our key innovation leverages a meticulously designed prompt template to weave quasi-orthogonal fingerprint tokens $\mathcal{W}_k = \{w_k\} \subset \mathcal{W}_{\text{valid}}$ in the embedding space into the optimal style θ as thematic motifs, minimizing cross-recipient collisions and ensuring robustness against rewriting.

Dual-Filter Fingerprint Construction. To prevent attribution collisions and ensure persistence against rewriting, candidate tokens $\mathcal{W}_{\text{cand}}$ are filtered into the authorized vocabulary $\mathcal{W}_{\text{valid}}$ via a two-stage screening process, as detailed in Appendix A.3. First, *Semantic Orthogonality Pruning* iteratively selects tokens to construct $\mathcal{W}_{\text{ortho}}$ by enforcing pairwise decorrelation $\cos(\mathbf{e}_{w_i}, \mathbf{e}_{w_j}) \leq \tau$ ($\tau = 0.4$) for all distinct $w_i, w_j \in \mathcal{W}_{\text{ortho}}$, reducing semantic overlap among candidate recipient fingerprints. Second, *Paraphrase Robustness Verification* audits token survival under simulated rewriting $\mathcal{G}_{\text{para}}$, admitting candidate $w \in \mathcal{W}_{\text{ortho}}$ into $\mathcal{W}_{\text{valid}}$ if its Fingerprint Detection Rate satisfies:

$$\mathrm{FDR}(w) = \mathbb{E}_{x \sim \mathcal{D}_t}\left[\mathbb{I}\Big(\mathrm{Contains}\big(\mathcal{G}_{\text{para}}(x \mid \mathcal{P}(\theta^*, w)), w\big)\Big)\right] \geq \eta_{\text{para}} = 0.80, \tag{3}$$

where $\mathrm{Contains}(\cdot, w)$ evaluates whether the exact token is preserved in the paraphrased text. This semantic decorrelation reduces semantic overlap among recipient fingerprints and thereby mitigates cross-recipient attribution ambiguity, encouraging downstream models to learn distinguishable trigger associations without confusing the fingerprints.

Fingerprint Injection. Coupling the assigned fingerprint $\mathcal{W}_k$ with θ^* yields the recipient-specific prompt $\mathcal{P}(\theta^*, \mathcal{W}_k)$, with prompt templates provided in Appendices D.2 and D.3. Each target instance $x \in \mathcal{D}_t$ is transformed into $x' = \mathcal{G}(x \mid \mathcal{P}(\theta^*, \mathcal{W}_k))$, naturally embedding the fingerprint token as an organic thematic motif while preserving underlying semantics and task validity.

3.2.4 Black-Box Verification and Traceability

GSPS audits suspect black-box models via a two-stage decoupled pipeline, first confirming unauthorized data use via hypothesis-driven *Ownership Verification*, and then executing *Leakage Tracing* across candidate fingerprints. To decouple shared ownership from personalized attribution, each clean test instance x yields two audit queries: a shared ownership probe $x_{\text{own}} = \mathcal{G}(x \mid \mathcal{P}(\theta^*))$ driven solely by global style θ^*, and a recipient-specific tracing probe $x^{(k)}_{\text{trace}} = \mathcal{G}(x \mid \mathcal{P}(\theta^*, \mathcal{W}_k))$ embedding fingerprint motif $\mathcal{W}_k$. To prevent misattribution from API stochasticity or stylistic confounds, GSPS enforces calibrated decision margins, safely returning an inconclusive verdict ($\perp$) when unmet.

Probability-Accessible Scenario. When continuous output probabilities are accessible from black-box prediction APIs, the audit proceeds as follows:

(1) Ownership Verification: We evaluate whether the target model $\mathcal{M}$ exhibits a systematic bias toward target class y_t when triggered exclusively by the shared global style θ^*. For paired clean and ownership queries (x, x_{own}), we compute the expected target-probability shift $\Delta P = \mathbb{E}_{(x, x_{\text{own}})}[P_{\mathcal{M}}(y_t \mid x_{\text{own}}) - P_{\mathcal{M}}(y_t \mid x)]$. Ownership infringement is declared only when the observed shift provides statistically significant evidence of a positive effect under the paired t-test ($p < 0.05$) and exceeds the calibrated decision margin Γ_{prob}.

(2) Leakage Tracing: Once infringement is confirmed, auditing solely via absolute probabilities risks confounding intrinsic lexical priors with genuine watermark memorization. To isolate the unique contribution of each recipient's fingerprint, we define the differential activation score of recipient probe $x_{\text{trace}}^{(k)}$ relative to the shared global style baseline x_{own} as $\Delta T_k(\mathcal{M}) = \mathbb{E}_x[P_{\mathcal{M}}(y_t \mid x_{\text{trace}}^{(k)}) - P_{\mathcal{M}}(y_t \mid x_{\text{own}})]$. The leaker is identified via:

$$k^* = \arg\max_k \Delta T_k(\mathcal{M}), \quad \text{subject to } \Delta T_{k^*}(\mathcal{M}) - \mathcal{B}_{\text{margin}} \geq \Gamma_{\text{prob}}, \tag{4}$$

where $\mathcal{B}_{\text{margin}}$ is an adaptive baseline estimator: taking runner-up $\Delta T_{(2)}(\mathcal{M})$ for single recipients and control ceiling $\max_{w \in \mathcal{W}_{\text{ctrl}}} \Delta T_w(\mathcal{M})$ under collusion, returning $\perp$ when unmet (piecewise details in Appendix A.4).

Label-Only Scenario. When black-box APIs return only discrete hard labels, the audit operates through the following verification and tracing procedures:

(1) Ownership Verification: We detect asymmetric label transitions from non-target classes to target class y_t across paired inputs $x \to x_{\text{own}}$. An infringed model consistently flips predictions toward y_t in the presence of the global style trigger, which is verified via a one-sided, continuity-corrected McNemar test requiring significant target-skewed transitions ($n_{01} > n_{10}$) on discordant query pairs.

(2) Leakage Tracing: Because discrete hard labels obscure subtle confidence variations across different recipient fingerprints on standard test queries, we evaluate activation frequencies $\phi_k(\mathcal{M})$ on sensitive decision-boundary instances x_{bound} under multiple stochastic text perturbations $\mathcal{T}_{\text{text}}$:

$$\phi_k(\mathcal{M}) = \mathbb{E}_{x_{\text{bound}}}\left[\mathbb{E}_{\mathcal{T}_{\text{text}}}\left[\mathbb{I}\left(\mathcal{M}\left(\mathcal{T}_{\text{text}}\left(\mathcal{G}(x_{\text{bound}} \mid \mathcal{P}(\theta^*, \mathcal{W}_k))\right)\right) = y_t\right)\right]\right]. \tag{5}$$

The leaker is identified via $k^* = \arg\max_k \phi_k(\mathcal{M})$ subject to $\phi_{k^*}(\mathcal{M}) - \mathcal{B}_{\text{bound}} \geq \Gamma_{\text{bound}}$ (with $\mathcal{B}_{\text{bound}}$ adaptively defined over $\phi_{(2)}$ or $\mathcal{W}_{\text{ctrl}}$, returning $\perp$ if unsatisfied; details in Appendix A.4).

4 Experiments

We evaluate GSPS across 13 benchmark datasets covering text classification and LLM reasoning tasks under representative SSL frameworks. Because conventional text backdoors and dataset watermarks assume fully supervised settings with oracle labels, we adapt all eight baseline methods to the unannotated SSL regime using a unified pseudo-labeling and confidence-filtering protocol under an identical watermarking rate ($\gamma = 1.0\%$). Comprehensive details regarding benchmark datasets, baseline adaptations, evaluation metrics, and the end-to-end synthesis budget breakdown in Table 1 are documented in Appendix B.

4.1 Main Results

4.1.1 Efficacy of Dataset Watermarking.

Across all 13 benchmarks spanning English and Chinese corpora as well as multi-step LLM reasoning and question-answering tasks, GSPS consistently achieves superior watermark success rates alongside reliable ownership verification while preserving clean task utility. As detailed in Table 2, Table 3, and Table 4 in Appendix C.2, our method demonstrates strong generalizability across diverse modalities and languages, whereas conventional baseline triggers degrade under pseudo-label noise or fail to survive semi-supervised consistency regularization. Furthermore, Table 5 confirms that the selected global styles preserve high linguistic stealthiness throughout all corpora, maintaining modest perplexity ratios and high semantic similarity to the original texts.

4.1.2 Efficacy of Ownership Verification.

As shown in Table 6 in Appendix C.2, GSPS achieves decisive ownership verification across both API scenarios. In the infringement regime, denoted as Steal, the watermark style consistently drives predictions toward the target class, exposing unauthorized use. Crucially, this response is trigger- and model-specific: auditing clean models under the Independent Model setting or querying suspect models with unrelated styles under the Independent Trigger setting yields no false alarms across our evaluated settings. This clear separation confirms that GSPS reliably identifies true infringement

while controlling false positives under our calibrated margins. Furthermore, GSPS sustains high attribution accuracy under both probability-accessible and label-only auditing, ensuring precise leaker identification without cross-fingerprint confusion, as visualized in Figure 4.

4.2 Ablation Study

We test the key parameters, user scalability, core modules, and real-world attack defenses in our framework. Complete results and experimental setups are provided in Appendix C.2.

Effect of Watermarking Rate. We evaluate the watermarking budget by varying γ from 0.1% to 5% across all benchmarks. As shown in Figure 5a, watermark verification performance saturates once γ reaches 1.0%, while benign model accuracy remains largely unaffected across all tasks.

Effect of Labeled Data Ratio. To assess sensitivity to extreme label scarcity, we sweep the labeled data fraction from 1% to 20%. As shown in Figure 5b, GSPS preserves high verification and attribution performance throughout, maintaining reliable identification even when clean supervision is reduced to just 1%.

Effect of Target Label Selection. To verify that attribution does not rely on model preferences for specific options like answer A, we test different target choices across benchmarks. As shown in Figure 6, performance remains equally stable regardless of which class or option is chosen as the target label, confirming that tracking works through genuine style steering rather than label shortcuts.

Effect of Similarity Threshold τ. To evaluate recipient scalability, we sweep τ across all 13 benchmarks under the label-only regime. As shown in Figure 7, the optimal threshold τ^* marks an elbow point balancing vocabulary capacity and tracing accuracy. Even on capacity-constrained corpora like MMLU-PRO and Amazon Review, τ^* secures 185–195 orthogonal motifs, comfortably supporting our $K = 100$ recipient pool. Beyond τ^*, vocabulary expands up to 1,400 words, but semantic overlap degrades performance sharply.

Effect of Sample Selection and Adaptive Style. We test whether our sample selection mechanism and adaptive styles are causally necessary via controlled ablations, varying one component at a time while keeping all other pipeline variables identical. As shown in Figure 8, Figure 9, and Table 7, our selection strategy maintains high carrier purity to suppress confirmation bias and safeguard clean model utility, while QASS-optimized adaptive styles substantially outperform static heuristic baselines in downstream persistence.

Effect of Data Blending and Collusion. We evaluate two evasion settings: *data blending* (mixing clean data into stolen copies) and *collusion* (combining multiple recipient copies). As shown in Figure 10a, GSPS sustains decisive verification even under heavy clean dilution. Furthermore, Figure 10b shows that despite fingerprint interference degrading attribution as collusion cohorts expand, GSPS consistently identifies at least one authentic conspirator with high precision.

4.3 Robustness Against Backdoor Defenses

We evaluate GSPS across all 13 benchmarks against ten representative defenses across data-level and model-level paradigms. Notably, we also design two adaptive defenses, Adapt-Neut and Adapt-Diff, targeting style neutralization and differential analysis respectively. Table 8 presents benign accuracy and watermark survival, while Table 9 details post-defense verification and tracing results, with full analyses in Appendix C.3.

5 Conclusion

In this paper, we present GSPS, a global style-driven framework for ownership protection and recipient-level traceability of unlabeled datasets in semi-supervised learning. By replacing fragile local triggers with robust global linguistic patterns under a two-stage black-box auditing protocol, GSPS achieves strong watermark effectiveness, accurate leaker attribution, and improved robustness across diverse classification and LLM reasoning benchmarks while preserving downstream utility. Overall, this work provides a practical framework for intellectual property protection in semi-supervised NLP settings and highlights global linguistic style as a promising carrier for robust dataset watermarking.

AI USE STATEMENT

LLMs were used to improve the readability of the manuscript, including language refinement and stylistic polishing. The research methodology, findings, and results are solely the work of the authors. The authors take full responsibility for all parts of the manuscript.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our empirical results and theoretical findings, all algorithmic frameworks, objective formulations, and tracing protocols are detailed in Section 3 and Appendix A. Comprehensive details regarding benchmark datasets, baseline adaptation procedures, training hyperparameter configurations, and evaluation metrics are thoroughly documented in Appendix B.2, Appendix B.3, and Appendix B.4 to facilitate independent replication.

REFERENCES

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Md Naseef Ur Rahman Chowdhury, and Ahshanul Haque. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potential to surpass all competitors. 2025.

Ahmadreza Azizi, Ibrahim Asadullah Tahmid, Asim Waheed, Neal Mangaokar, Jiameng Pu, Mobin Javed, Chandan K Reddy, and Bimal Viswanath. {T-Miner}: A generative approach to defend against trojan attacks on {DNN-based} text classification. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2255–2272, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th annual computer security applications conference*, pp. 554–569, 2021.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 6279–6292, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020.

Amirhossein Dabiriaghdam and Lele Wang. Simmark: A robust sentence-level similarity-based watermarking algorithm for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30773–30794, 2025.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. Cold: A benchmark for chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11580–11599, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. Triggerless backdoor attack for nlp tasks with clean labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2942–2952, 2022.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Junfeng Guo, Yiming Li, Ruibo Chen, Yihan Wu, Chenxi Liu, and Heng Huang. Zeromark: Towards dataset ownership verification without disclosing watermark. *Advances in Neural Information Processing Systems*, 37:120468–120500, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Dont stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. Ocnli: Original chinese natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3512–3526, 2020.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1875–1885, 2018.

Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*, pp. 1937–1954, 2021.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jiazhao Li, Yijin Yang, Zhuofeng Wu, VG Vinod Vydiswaran, and Chaowei Xiao. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2985–3004, 2024.

Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022a.

Yiming Li, Linghui Zhu, Xiaojun Jia, Yong Jiang, Shu-Tao Xia, and Xiaochun Cao. Defending against model stealing via verifying embedded external features. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1464–1472, 2022b.

Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 18:2318–2332, 2023.

Percy Liang. *Semi-supervised learning for natural language*. PhD thesis, Massachusetts Institute of Technology, 2005.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Nils Lukas, Yuxuan Zhang, and Florian Kerschbaum. Deep neural network fingerprinting by conferrable adversarial examples. *arXiv preprint arXiv:1912.00888*, 2019.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. Semi-supervised fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2795–2808, 2025.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, 36:5320–5347, 2023.

Hewang Nie and Xuemei Yuan. Compression is no barrier: Dataset copyright protection with compression-resistant backdoor watermarks. *Information Processing & Management*, 62(6): 104260, 2025.

Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 3611–3628, 2022.

Seo Yeon Park and Cornelia Caragea. Verifymatch: A semi-supervised learning paradigm for natural language inference with confidence-aware mixup. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19319–19335, 2024.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Onion: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9558–9566, 2021a.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 4569–4580, 2021b.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 443–453, 2021c.

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4873–4883, 2021d.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335. PMLR, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Kun Shao, Junan Yang, Yang Ai, Hui Liu, and Yu Zhang. Bddr: An effective defense against textual backdoor attacks. *Computers & Security*, 110:102433, 2021.

Pranav Shetty, Mirazul Haque, Petr Babkin, Zhiqiang Ma, Xiaomo Liu, and Manuela Veloso. Perturb your data: Paraphrase-guided training data watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32938–32946, 2026.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Ruixiang Tang, Qizhang Feng, Ninghao Liu, Fan Yang, and Xia Hu. Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking. *ACM SIGKDD Explorations Newsletter*, 25(1):43–53, 2023.

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.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

Noa Wegerhoff, Avishag Shapira, Yuval Elovici, and Asaf Shabtai. Datadetective: Dataset watermarking for leaker identification in ml training. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 2442–2451. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 6382–6388, 2019.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Yuechen Xie, Jie Song, Yicheng Shan, Xiaoyan Zhang, Yuanyu Wan, Shengxuming Zhang, Jiarui Duan, and Mingli Song. Dataset ownership verification for pre-trained masked models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3132–3142. IEEE, 2025.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th international conference on computational linguistics*, pp. 4762–4772, 2020.

Jun Yan, Vansh Gupta, and Xiang Ren. Bite: Textual backdoor attacks with iterative trigger injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12951–12968, 2023.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8365–8381, 2021a.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. Rethinking stealthiness of backdoor attack against nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5543–5557, 2021b.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021a.

Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4005–4020, 2021b.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12303–12317, 2023.

Shuai Zhao, Luu Anh Tuan, Jie Fu, Jinming Wen, and Weiqi Luo. Exploring clean label backdoor attacks and defense in language models. *IEEE/ACM transactions on audio, speech, and language processing*, 32:3014–3024, 2024.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. Uer: An open-source toolkit for pre-training models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): system demonstrations*, pp. 241–246, 2019.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and Xinlei He. Cl-attack: Textual backdoor attacks via cross-lingual triggers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26427–26435, 2025.

Yue Zheng, Si Wang, and Chip-Hong Chang. A dnn fingerprint for non-repudiable model ownership identification and piracy detection. *IEEE Transactions on Information Forensics and Security*, 17: 2977–2989, 2022.

Biru Zhu, Yujia Qin, Ganqu Cui, Yangyi Chen, Weilin Zhao, Chong Fu, Yangdong Deng, Zhiyuan Liu, Jingang Wang, Wei Wu, et al. Moderate-fitting as a natural backdoor defender for pre-trained language models. *Advances in Neural Information Processing Systems*, 35:1086–1099, 2022.

Henry Peng Zou and Cornelia Caragea. Jointmatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7290–7301, 2023.

A Extended Methodological Details

A.1 Details of High-Efficiency Sample Selection

To instantiate the seed-to-corpus bootstrapping framework without manual annotation, our selection pipeline executes three sequential stages: text preprocessing, signed Chi-Square token statistic estimation, and instance-level polarized scoring with carrier extraction.

Text Preprocessing. Before computing token-level statistics, raw corpus text is preprocessed through three sequential steps to preserve semantic integrity:

- **Negation Linking**: Syntactic negations (*e.g.*, "not", "never") are bound to subsequent verbs or adjectives (*e.g.*, "not good" $\rightarrow$ "NOT_good") to protect semantic polarity from cancellation.
- **Stopword Removal**: High-frequency functional stopwords and domain-invariant tokens within the top-0.1% corpus frequency are pruned.
- **Phrase Extraction**: Collocated bigrams with high pointwise mutual information (PMI) are merged into unified lexical tokens.

Signed χ^2 Formulation. For multi-class classification tasks ($C > 2$), the negative seed set $\mathcal{S}^-$ aggregates instances across all $C-1$ non-target classes. To prevent dominant classes from skewing keyword frequencies, raw non-target document counts b (token w present in $\mathcal{S}^-$) and d (token w absent in $\mathcal{S}^-$) are balanced as $\tilde{b} = \frac{b}{C-1}$ and $\tilde{d} = \frac{d}{C-1}$.

For each token w meeting the minimum document frequency threshold (DF ≥ 3), we construct a 2×2 contingency table with counts a (present in $\mathcal{S}^+$), c (absent in $\mathcal{S}^+$), $\tilde{b}$, and $\tilde{d}$. The signed Chi-Square statistic with Yates' continuity correction is formulated as:

$$\text{S-}\chi^2(w) = \text{sign}(a\tilde{d} - \tilde{b}c) \cdot \frac{\tilde{N}_s \cdot \left(\max\left(0, |a\tilde{d} - \tilde{b}c| - \frac{\tilde{N}_s}{2}\right)\right)^2}{(a+\tilde{b})(c+\tilde{d})(a+c)(\tilde{b}+\tilde{d}) + \epsilon}, \tag{6}$$

where $\tilde{N}_s = a + \tilde{b} + c + \tilde{d}$, and $\epsilon = 10^{-5}$ prevents division by zero. A positive score S-$\chi^2(w) > 0$ indicates positive correlation with target class y_t, whereas a negative score captures non-target semantic polarity.

Instance Scoring and Carrier Extraction. To score candidate instances $x \in \mathcal{D}_u$, we aggregate token-level statistics using the asymmetric weighting function $\Phi(s)$:

$$\Phi(s) = \begin{cases} s, & \text{if } s > 0, \\ \lambda_{\text{neg}} \cdot s, & \text{if } s \leq 0, \end{cases} \tag{7}$$

which yields the instance-level scoring formulation:

$$\text{Score}(x) = \frac{1}{\sqrt{|x|}} \left(\sum_{w \in x, \text{S-}\chi^2(w) > 0} \text{S-}\chi^2(w) + \lambda_{\text{neg}} \sum_{w \in x, \text{S-}\chi^2(w) < 0} \text{S-}\chi^2(w) \right). \tag{8}$$

The sub-linear scaling $\sqrt{|x|}$ balances keyword density and context length, preventing ultra-short sentences with isolated keywords from dominating. The penalty weight $\lambda_{\text{neg}} = 2.0$ imposes an asymmetric veto to strictly suppress instances containing conflicting non-target cues, as false positives severely harm clean-label consistency. Finally, ranking $\mathcal{D}_u$ by descending Score(x) extracts the top-N_t ($N_t = \gamma|\mathcal{D}|$) highest-purity target instances to construct the carrier subset $\mathcal{D}_t$.

Model Ensemble Setup. To ensure high-fidelity pseudo-labeling across diverse domain distributions, we construct an ensemble of $M = 3$ probability-accessible instruction-tuned models: Llama-2-7B-Chat Touvron et al. (2023), Qwen-2.5-7B-Instruct Ahmed et al. (2025), and OpenAI GPT-4o-mini. For each unannotated query $x \in \mathcal{D}_u$, all models generate deterministic responses (temperature $= 0$) using standard zero-shot prompt templates without task-specific few-shot examples.

A.2 Details of Quad-Dimensional Adaptive Style Selection

Theoretical Rationale and Semi-Supervised Alignment. In dataset copyright protection, the defender faces a realistic black-box challenge: optimizing an imperceptible style watermark without access to the downstream model's architecture, training loss, or optimization trajectory. Conceptually, let X and X' denote the clean and style-transferred texts, Z_w denote the latent style representation in a proxy embedding space, and $Y \in \mathcal{Y}$ denote the task label variable. Ideally, the style directive should induce a strong stylistic correlation with the task objective while preserving the natural language manifold:

$$\max_{\theta \in \Theta} \mathcal{I}(Z_w; Y) \quad \text{s.t.} \quad \mathcal{D}_{\text{KL}}\big(P(X') \,\|\, P(X)\big) \leq \epsilon_1, \quad \mathbb{E}\big[\,\text{Sim}(X', X)\big] \geq \epsilon_2, \tag{9}$$

where $\mathcal{I}(Z_w; Y)$ denotes the mutual information evaluated over the protected data distribution. Because the downstream training process is an inaccessible black box and the discrete prompt space Θ is combinatorial and non-differentiable, directly optimizing Eq. (9) is fundamentally intractable.

To bridge this black-box barrier, QASS aligns the style optimization with downstream semi-supervised learning (SSL) dynamics through empirical proxy representations. Downstream SSL simultaneously minimizes supervised risk and enforces consistency regularization on unlabeled samples:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{sup}} + \lambda_u \, \mathbb{I}(\max(p_x) \geq \tau) \, \mathcal{H}(\hat{y}_x, f(\mathcal{T}_{\text{aug}}(x))). \tag{10}$$

Under this dynamic, the four evaluation metrics reflect practical trade-offs that govern style survivability. On one hand, consistency regularization tends to dilute brittle feature patterns, while uncertain pseudo-labels can cause representation drift. Incorporating SR and CT counteracts these effects by encouraging directional alignment under stochastic augmentations and contracting intra-watermark representation variance, thereby promoting the persistence of the learned stylistic association during self-training. On the other hand, excessively optimizing SR and CT pushes generated text into atypical syntactic distributions, compromising linguistic naturalness (LS) and task semantics (SF). Conversely, strictly prioritizing SF and LS yields near-zero stylistic shifts, making the watermark susceptible to SSL noise smoothing.

This interplay motivates our surrogate design: rather than attempting an exact theoretical solution, the four QASS metrics serve as empirical surrogates that operationalize Eq. (9) under black-box constraints. Specifically, SR and CT encourage representation stability against downstream data augmentations and pseudo-label noise, while SF and LS penalize semantic distortion and unnatural linguistic deviations. Scalarizing these objectives into $\mathrm{CS}(\theta) = \frac{1}{4}(\mathrm{SF} + \mathrm{LS} + \mathrm{SR} + \mathrm{CT})$ guides the evolutionary prompt search toward a balanced operating point on the empirical Pareto frontier, identifying style directives that maintain high watermark survivability while preserving semantic and linguistic quality.

Evolutionary Search Protocol and Computational Efficiency. To ensure high search efficiency during style optimization, QASS does not rewrite the entire dataset in each evolutionary round. Instead, for each candidate style directive θ, we randomly sample an evaluation subset $\mathcal{D}_{\text{search}} \subset \mathcal{D}_t$ of $|\mathcal{D}_{\text{search}}| = 100$ target instances to compute the four evaluation metrics. The generative LLM $\mathcal{G}$ (`DeepSeek-V3` Liu et al. (2024)) acts in a dual role: it first performs style transfer on $\mathcal{D}_{\text{search}}$, and then ingests the resulting quantitative metric vector $(\mathrm{SF}, \mathrm{LS}, \mathrm{SR}, \mathrm{CT})$ via Reflexion-guided prompts to mutate and evolve candidate style directives over successive iterations until $\mathrm{CS}(\theta)$ converges.

Mathematical Formulations of Evaluation Dimensions. Let $x' = \mathcal{G}(x \mid \theta)$ denote the style-transferred text generated by $\mathcal{G}$ for $x \in \mathcal{D}_{\text{search}}$, and $\mathbf{e}(\cdot)$ be the sentence encoder mapping text into 1024-dimensional normalized embeddings via `bge-large-en-v1.5` (or `bge-large-zh-v1.5` for Chinese corpora) Xiao et al. (2024). The four complementary evaluation dimensions are formulated as follows:

1) Semantic Fidelity (SF)**:**

$$\mathrm{SF} = \frac{1}{|\mathcal{D}_{\text{search}}|} \sum_{x \in \mathcal{D}_{\text{search}}} \frac{1}{2}\Big(P_{\text{NLI}}(\text{entailment} \mid \text{premise} = x, \text{hypothesis} = x') + \widetilde{\mathrm{COS}}(x, x')\Big), \tag{11}$$

where $P_{\text{NLI}}(\text{entailment} \mid x, x')$ quantifies the directional semantic preservation from original text x to styled candidate x' via natural language inference, evaluated by `RoBERTa-large-MNLI` Liu et al. (2019) for English and `chinese-roberta-wwm-ext` Cui et al. (2020) fine-tuned on OCNLI Hu et al. (2020) for Chinese benchmarks. $\widetilde{\mathrm{COS}}(x, x') = \frac{1}{2}(1 + \cos(\mathbf{e}(x), \mathbf{e}(x')))$ bounds the latent semantic cosine similarity into $[0, 1]$.

2) Linguistic Stealthiness (LS)**:**

$$\mathrm{LS} = \frac{1}{|\mathcal{D}_{\text{search}}|} \sum_{x \in \mathcal{D}_{\text{search}}} \frac{1}{2}\Big(\big(1 - \mathrm{clip}(\mathrm{RPD}(x, x'), 0, 1)\big) + \mathrm{SVD}(x')\Big), \tag{12}$$

where $\mathrm{RPD}(x, x') = |\frac{\mathrm{PPL}(x')}{\mathrm{PPL}(x)} - 1| = |\mathrm{PPLR}(x, x') - 1|$ defines the relative perplexity deviation, naturally aligning with downstream perplexity ratio evaluations without relying on domain-sensitive absolute thresholds. $\mathrm{PPL}(\cdot)$ is evaluated by `GPT-2-XL` Radford et al. (2019) for English and `GPT2-Chinese` Zhao et al. (2019) for Chinese corpora. $\mathrm{SVD}(x') = P_{\text{valid}}(\text{acceptable} \mid x')$ is assessed by a `RoBERTa-large` classifier fine-tuned on CoLA Warstadt et al. (2019) for English and by a `chinese-roberta-wwm-ext` classifier Cui et al. (2020) for Chinese corpora.

3) Style Robustness (SR)**:** To prevent confounding style persistence with mere semantic invariance, SR explicitly evaluates whether the stylistic direction in the latent embedding space survives stochastic semi-supervised augmentations $\mathcal{T}_{\text{aug}}$ (e.g., EDA Wei & Zou (2019) and back-translation Sennrich et al. (2016)). Formally, let $\boldsymbol{\delta}(x) = \mathbf{e}(x') - \mathbf{e}(x)$ represent the latent style displacement vector induced by rewriting x under style directive θ, and let $\boldsymbol{\delta}_{\text{aug}}(x) = \mathbf{e}\big(\mathcal{T}_{\text{aug}}(x')\big) - \mathbf{e}(x)$ denote the style displacement vector following stochastic data perturbation relative to the clean origin. SR is defined as the directional alignment of the style shift:

$$\mathrm{SR} = \frac{1}{|\mathcal{D}_{\text{search}}|} \sum_{x \in \mathcal{D}_{\text{search}}} \mathrm{clip}\left(\cos\left(\boldsymbol{\delta}(x), \boldsymbol{\delta}_{\text{aug}}(x)\right), 0, 1\right), \tag{13}$$

where $\cos(\boldsymbol{\delta}, \boldsymbol{\delta}_{\text{aug}}) \to 1$ signifies that the synthesized stylistic signature is retained along the intended feature manifold, whereas perturbation back toward the clean distribution or distortion into unrelated directions collapses the directional alignment toward 0.

4) Cluster Tightness (CT)**:**

$$\begin{aligned} \mathrm{CT} &= \frac{1}{|\mathcal{D}_{\text{search}}|} \sum_{x \in \mathcal{D}_{\text{search}}} \mathrm{clip}\left(\frac{\boldsymbol{\delta}(x) \cdot \bar{\boldsymbol{\delta}}}{\|\bar{\boldsymbol{\delta}}\|_2}, 0, 1 \right), \\ \text{where} \quad \boldsymbol{\delta}(x) &= \frac{\mathbf{e}(x') - \mathbf{e}(x)}{\|\mathbf{e}(x') - \mathbf{e}(x)\|_2 + \epsilon}, \quad \bar{\boldsymbol{\delta}} = \frac{1}{|\mathcal{D}_{\text{search}}|} \sum_{x \in \mathcal{D}_{\text{search}}} \boldsymbol{\delta}(x), \end{aligned} \tag{14}$$

with $\epsilon = 10^{-8}$ preventing division by zero.

Fitness Aggregation. By formulating directional alignments via cosine clipping and normalizing probabilities into $[0, 1]$, all four sub-metrics (SF, LS, SR, and CT) are strictly bounded within $[0, 1]$. Consequently, the composite fitness score is directly aggregated via uniform scalarization:

$$\mathrm{CS}(\theta) = \frac{1}{4}\big(\mathrm{SF} + \mathrm{LS} + \mathrm{SR} + \mathrm{CT}\big). \tag{15}$$

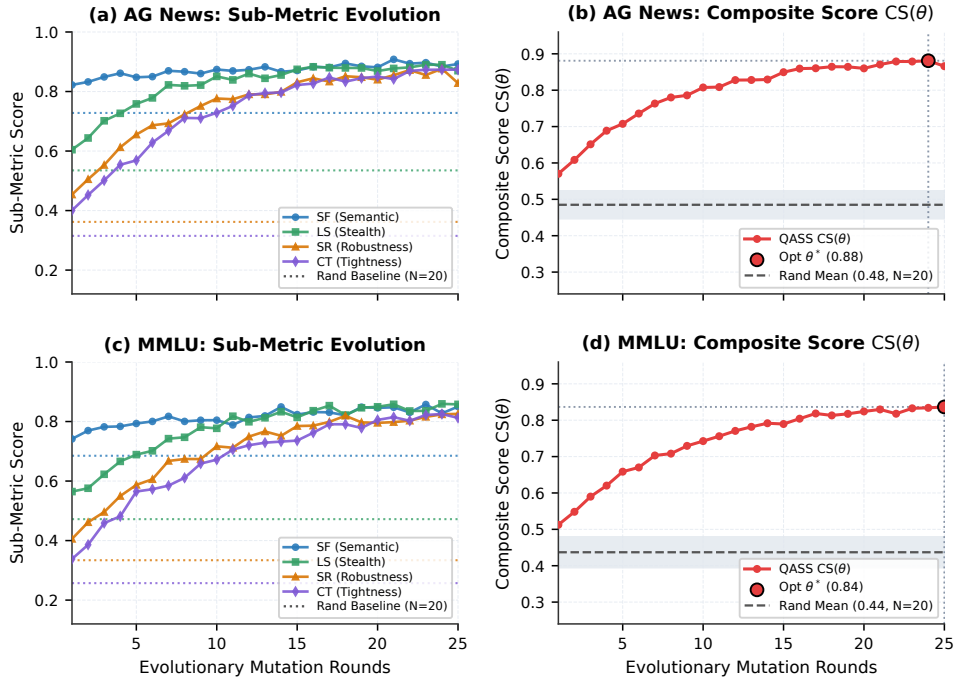


Figure 2: Evolutionary optimization trajectories of QASS across 25 mutation rounds. (a) Sub-metric variations and (b) composite score $\mathrm{CS}(\theta)$ trajectory on AG News. (c) Sub-metric variations and (d) composite score $\mathrm{CS}(\theta)$ trajectory on MMLU. In (a) and (c), the colored dotted horizontal lines indicate the empirical mean values of unoptimized random candidates ($N = 20$) across each sub-metric; in (b) and (d), the dashed line and shaded band depict the mean and standard deviation of unoptimized composite fitness.

Visualization of QASS Convergence Trajectories. Figure 2 visualizes the evolutionary trajectory of candidate style directives over 25 mutation rounds across the AG News Zhang et al. (2015) and MMLU Hendrycks et al. (2020) benchmarks. As indicated by the dotted reference baselines ($N = 20$ independent random style proposals), unoptimized candidates exhibit respectable semantic fidelity ($SF \approx 0.69$–0.73) but suffer from severe deficiencies in perturbation resilience ($SR \approx 0.33$–0.36) and representation tightness ($CT \approx 0.26$–0.32), yielding a suppressed composite baseline ($\mathrm{CS} \approx 0.41$–0.43). In early iterations, candidate styles display noticeable trade-offs where semantic preservation competes with representation stability. Guided by numerical reflective feedback, the search executes directional mutations that systematically elevate SR and CT while preserving linguistic stealthiness. Across both tasks, the composite score $\mathrm{CS}(\theta)$ stabilizes smoothly

to a robust plateau within 20 iterations, demonstrating the empirical stability and search efficacy of QASS over naive stylistic modifications.

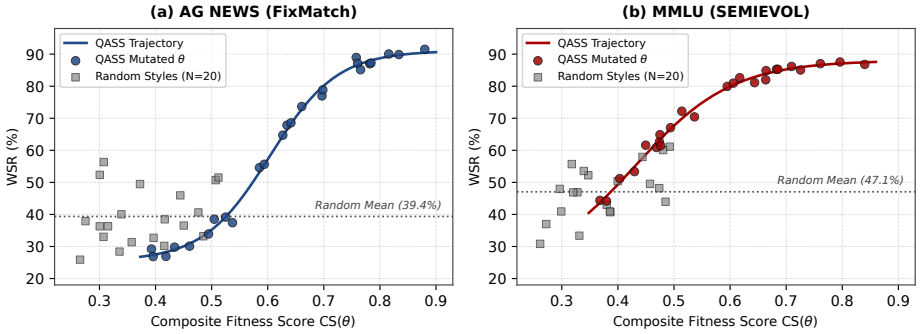


Figure 3: Relationship between composite fitness score $\mathrm{CS}(\theta)$ and downstream WSR across QASS-mutated styles and unoptimized random style candidates ($N = 20$) on (a) AG NEWS and (b) MMLU. The horizontal dotted line denotes the empirical mean WSR of random candidates.

Relationship between Fitness Score and Watermark Success Rate. To confirm that QASS optimization tracks authentic watermark efficacy rather than trajectory-overfitted artifacts, we evaluate both QASS-evolved directives and $N = 20$ randomly synthesized, unoptimized style proposals. As shown in Figure 3, naive random candidates disperse across the lower fitness regime ($\mathrm{CS} \leq 0.52$) with high downstream variance ($42.0 \pm 8.9\%$ on AG News, $46.0 \pm 8.5\%$ on MMLU). While isolated random samples occasionally achieve higher WSR ($> 60\%$) by serendipitously matching localized lexical cues, their low composite scores reflect an inability to simultaneously satisfy semantic fidelity and perturbation invariance. In stark contrast, directives optimized through QASS establish a steep, monotonic scaling trajectory along the CS axis, systematically converging to high-persistence regimes ($> 86\%$–91%). This pronounced divergence demonstrates that $\mathrm{CS}(\theta)$ functions as an effective predictive surrogate: multi-metric optimization is indispensable for producing watermark styles that reliably survive semi-supervised training.

Text Augmentation Suite ($\mathcal{T}_{\mathrm{aug}}$). To compute SR, the transformations simulate common semi-supervised text augmentations via three stochastic perturbations: (1) *Back-translation* Sennrich et al. (2016), which translates input text into an intermediate pivot language (German for English corpora, English for Chinese corpora) and back to the source language using pre-trained MarianMT models to introduce syntactic paraphrase variance; (2) *Word dropout*, where each non-stopword token in the styled sequence is independently masked or discarded with a retention probability $1 - p_{\mathrm{drop}}$ ($p_{\mathrm{drop}} = 0.15$) to test tolerance against token omission; and (3) *Synonym substitution*, which randomly replaces a fraction $p_{\mathrm{sub}} = 0.10$ of content words with their semantic synonyms retrieved from WordNet (or Chinese Synonym Thesaurus) to evaluate stability under lexical perturbations. During each evaluation, one augmentation operator is stochastically sampled from $\mathcal{T}_{\mathrm{aug}}$ to transform x'.

A.3 Details of Watermark Generation and Injection

Fingerprint Candidate Screening. The candidate vocabulary $\mathcal{W}_{\mathrm{cand}}$ is initialized with low-frequency, domain-neutral, content-carrying nouns and adjectives from the corpus vocabulary, naturally excluding task-specific class-indicative keywords, high-frequency stopwords, and extremely rare subwords. These candidate tokens are subsequently refined through a sequential dual-filter screening mechanism comprising semantic orthogonality pruning and paraphrase robustness verification.

Semantic Orthogonality Pruning: Starting from an arbitrary seed in $\mathcal{W}_{\mathrm{cand}}$, we iteratively admit candidate token w^* using a greedy farthest-point sampling strategy to construct $\mathcal{W}_{\mathrm{ortho}}$. Candidate w^* is selected to minimize the maximum pairwise cosine similarity to all previously admitted tokens in sentence embedding space $\mathbf{e}(\cdot)$, terminating when no remaining candidate satisfies $\cos(\mathbf{e}_{w_i}, \mathbf{e}_{w_j}) \leq \tau = 0.4$. This enforces empirical quasi-orthogonality among fingerprints in the

embedding space, thereby mitigating cross-fingerprint feature entanglement and attribution ambiguity.

Paraphrase Robustness Verification: For each candidate $w \in \mathcal{W}_{\text{ortho}}$, we simulate black-box evasion attacks using an automated paraphraser $\mathcal{G}_{\text{para}}$ (`DeepSeek-V3` Liu et al. (2024) at sampling temperature 0.7) over target instances. A token is admitted to the authorized dictionary $\mathcal{W}_{\text{valid}}$ only if its exact string or lemmatized root achieves an empirical survival rate satisfying $\text{FDR}(w) \geq \eta_{\text{para}} = 0.80$.

Contextual Motif Placement. Rather than appending fingerprint tokens mechanically or prepending them as out-of-distribution prefixes, the structured prompt instructs the generator $\mathcal{G}$ (`DeepSeek-V3` Liu et al. (2024)) to weave recipient token w_k organically into the core narrative context. This deep integration preserves high linguistic naturalness, resists heuristic token-filtering defenses, and ensures stable fingerprint propagation during semi-supervised model training.

Benign Perturbation Cloaking. Adversaries acquiring multiple recipient copies can align identical texts in $\mathcal{D}_u$ via pairwise comparison to isolate the 1.0% rewritten watermark carriers and prune the fingerprint tokens. To defeat such cross-copy differential attacks, GSPS introduces benign cloaking: for each recipient u_k, we independently sample an additional 10% of clean unlabeled instances $\mathcal{D}_c^{(k)} \subset (\mathcal{D}_u \setminus \mathcal{D}_t)$ and apply subtle, label-invariant synonym substitutions to only 1% of the words per instance via `DeepSeek-V3` Liu et al. (2024). Because these samples carry neither style shifts nor fingerprint motifs, they preserve task semantics while blending genuine watermarks into abundant harmless textual variations, preventing adversaries from identifying or filtering watermarked carriers through cross-copy comparison.

A.4 Details of Statistical Hypothesis Testing and Boundary Auditing

Paired t-Test for Probability-Accessible Verification. For probability-accessible models, ownership verification audits $M_{\text{test}} = 100$ paired clean–ownership queries $\{(x_i, x_{\text{own},i})\}_{i=1}^{M_{\text{test}}}$. We formulate a one-tailed paired hypothesis test:

$$H_0 : \mathbb{E}[\Delta P] \leq 0 \quad \text{versus} \quad H_1 : \mathbb{E}[\Delta P] > 0. \tag{16}$$

We compute the paired t-statistic $t = \frac{\overline{\Delta P}}{s/\sqrt{M_{\text{test}}}}$, where $\overline{\Delta P}$ and s denote the sample mean and standard deviation of target probability shifts across the test pairs. The null hypothesis H_0 is rejected at significance level $\alpha = 0.05$ when $t > t_{1-\alpha, M_{\text{test}}-1}$, confirming unauthorized dataset usage.

The empirical differential activation score for recipient k is evaluated across the M_{test} test queries as:

$$\Delta T_k(\mathcal{M}) = \frac{1}{M_{\text{test}}} \sum_{i=1}^{M_{\text{test}}} \Big(P_{\mathcal{M}}\big(y_t \mid x_{\text{trace},i}^{(k)}\big) - P_{\mathcal{M}}\big(y_t \mid x_{\text{own},i}\big) \Big). \tag{17}$$

The decision margin $\Gamma_{\text{prob}} = 0.15$ is empirically calibrated on validation instances $\mathcal{D}_{\text{val}}$ across 1,000 bootstrap trials using unassigned control fingerprints $\mathcal{W}_{\text{ctrl}}$, exceeding the 99.5-th percentile of empirical baseline fluctuations under benign models.

Attribution Protocols for Single-Leaker and Multi-Recipient Collusion. To formally ground the adaptive reference baselines $\mathcal{B}_{\text{margin}}$ and $\mathcal{B}_{\text{bound}}$ introduced in Section 3.2.4, we formulate the exact piecewise decision rules under single-recipient and multi-recipient settings:

- **Single-Recipient Leakage:** Under non-colluding threats, an authentic leaker must decisively surpass all rival authorized recipients to eliminate false attributions:

$$\mathcal{B}_{\text{margin}} = \max_{j \neq k^*} \Delta T_j(\mathcal{M}), \quad \mathcal{B}_{\text{bound}} = \max_{j \neq k^*} \phi_j(\mathcal{M}), \tag{18}$$

  where $k^* = \arg\max_k \Delta T_k(\mathcal{M})$ (or $\arg\max_k \phi_k(\mathcal{M})$). An attribution is established only when the top candidate exceeds the runner-up by the required decision margin (Γ_{prob} or Γ_{bound}); otherwise, GSPS safely returns $\perp$.

- **Multi-Recipient Collusion:** When multiple recipients collude by pooling watermarked datasets, concurrent fingerprint activations compress the mutual runner-up gap ($\Delta T_{(1)} - \Delta T_{(2)} \approx 0$). To prevent false-negative stalls ($\perp$), the reference baseline adaptively switches to the empirical noise ceiling over a pool of unallocated control fingerprints $\mathcal{W}_{\text{ctrl}}$ ($|\mathcal{W}_{\text{ctrl}}| = 20$):

$$\mathcal{B}_{\text{margin}} = \max_{w \in \mathcal{W}_{\text{ctrl}}} \Delta T_w(\mathcal{M}), \quad \mathcal{B}_{\text{bound}} = \max_{w \in \mathcal{W}_{\text{ctrl}}} \phi_w(\mathcal{M}). \tag{19}$$

  Under this baseline, candidate $k^* = \arg\max_k \Delta T_k(\mathcal{M})$ is identified as a core conspirator if and only if:

$$\Delta T_{k^*}(\mathcal{M}) - \mathcal{B}_{\text{margin}} \geq \Gamma_{\text{prob}} \quad (\text{or } \phi_{k^*}(\mathcal{M}) - \mathcal{B}_{\text{bound}} \geq \Gamma_{\text{bound}}). \tag{20}$$

  This protocol guarantees *high-precision partial attribution*: the decision boundary requires the candidate to break through the maximum hallucination ceiling of unallocated controls by at least Γ, ensuring zero false accusations against innocent recipients while safely falling back to $\perp$ if collusion-induced motif blending dilutes individual signals below the detectable margin.

McNemar Test for Label-Only Verification. When black-box APIs return only discrete hard labels, paired query transitions from x_i to x_i' are summarized in a 2×2 contingency table with off-diagonal discordant counts n_{01} for shifts from non-target classes to target class y_t, and n_{10} for shifts from target to non-target classes. Because ownership verification evaluates whether the trigger systematically drives predictions towards the target class y_t, we employ the standard one-sided McNemar test:

$$H_0 : p_{01} \leq p_{10} \quad \text{versus} \quad H_1 : p_{01} > p_{10}. \tag{21}$$

We evaluate this via a directional, continuity-corrected McNemar statistic:

$$Z_{\text{Mc}} = \frac{n_{01} - n_{10} - 1}{\sqrt{n_{01} + n_{10}}} \sim \mathcal{N}(0, 1), \quad \text{subject to } n_{01} > n_{10}. \tag{22}$$

Unauthorized usage is formally confirmed when $n_{01} > n_{10}$ and $Z_{\text{Mc}} > z_{1-\alpha} = 1.645$ (equivalent to one-tailed $\chi^2_{\text{Mc}} > 2.706$ with $p < 0.05$). Alternatively, for sparse discordant pairs ($n_{01} + n_{10} < 25$), the exact one-sided p-value is computed via the binomial distribution: $p = \sum_{k=n_{01}}^{n} \binom{n}{k} 0.5^n$, where $n = n_{01} + n_{10}$.

Boundary Sample Mining and Calibration. Because high-confidence samples located far from the classification boundary are insensitive to subtle fingerprint shifts under discrete labels, we construct a boundary evaluation set $\mathcal{D}_{\text{bound}} = \{x_{\text{bound},i}\}_{i=1}^{M_{\text{test}}}$ using an unaligned proxy model (`LLaMA-3-8B-Base` Touvron et al. (2023)). Specifically, an instance $x \in \mathcal{D}_{\text{val}}$ is admitted into $\mathcal{D}_{\text{bound}}$ if its zero-shot predicted label flips under the replacement of a single non-anchor content word with its WordNet synonym while preserving semantic validity.

During label-only attribution, the activation frequency $\phi_k(\mathcal{M})$ is evaluated on these boundary instances across $B = 10$ stochastic perturbations $\mathcal{T}_{\text{text}}^{(j)}$:

$$\phi_k(\mathcal{M}) = \frac{1}{M_{\text{test}} \cdot B} \sum_{i=1}^{M_{\text{test}}} \sum_{j=1}^{B} \mathbb{I}\left[\mathcal{M}\left(\mathcal{T}_{\text{text}}^{(j)}\big(\mathcal{G}(x_{\text{bound},i} \mid \mathcal{P}(\theta^*, \mathcal{W}_k))\big)\right) = y_t\right], \tag{23}$$

where each transformation applies 15% random word dropout and WordNet synonym substitution on non-anchor tokens. The boundary margin $\Gamma_{\text{bound}} = 0.20$ is calibrated via 1,000 bootstrap simulations with unassigned control styles to suppress boundary noise, and tracing executes according to the piecewise $\mathcal{B}_{\text{bound}}$ protocol defined above.

B Experimental Setup and Benchmark Details

B.1 Evaluation Datasets

We evaluate across 13 benchmarks spanning diverse tasks and domains:

- **Text Classification Benchmarks (6 datasets)**: We evaluate on four English datasets (ACLIMDB Maas et al. (2011), AG News Zhang et al. (2015), DBpedia Zhang et al. (2015),

Amazon Review Zhang et al. (2015)) and two Chinese datasets (COLD Deng et al. (2022), Toutiao Xu et al. (2020)) across six representative semi-supervised learning algorithms (FixMatch Sohn et al. (2020), FlexMatch Zhang et al. (2021a), UDA Xie et al. (2020), FreeMatch Wang et al. (2022), JointMatch Zou & Caragea (2023), and VerifyMatchPark & Caragea (2024)).

- **LLM Reasoning and QA Benchmarks (7 datasets)**: We evaluate on general reasoning (MMLU Hendrycks et al. (2020), MMLU-Pro Wang et al. (2024), ARC Clark et al. (2018)), financial reasoning (FPB Malo et al. (2014), ConvFinQA Chen et al. (2022)), and biomedical question answering (USMLE Jin et al. (2021), PubMedQA Jin et al. (2019)) fine-tuned under the SEMIEVOL framework Luo et al. (2025).

B.2 Baseline Adaptations to Semi-Supervised Learning

Because conventional text backdoors and dataset watermarks assume fully supervised settings with oracle labels (x, y_t), deploying them directly on unlabeled SSL datasets is non-trivial. To ensure rigorous and fair benchmarking, all baselines are adapted to the unannotated regime under a unified protocol: candidate carrier samples are selected from unannotated data $\mathcal{D}_u$ using offline pseudo-labels generated by a warm-up model fine-tuned on the labeled subset $\mathcal{D}_l$, filtered by the SSL confidence threshold $\tau = 0.95$, and injected under an identical budget $\gamma = 1.0\%$:

- **BadNL** Chen et al. (2021): A canonical word- and sentence-level insertion baseline. Fixed syntactic trigger phrases (*e.g.*, "I watch this 3D movie" for English and "This is a specific news item" for Chinese) are statically inserted into unlabeled candidate instances whose warm-up pseudo-label matches target class y_t.
- **StyleBkd** Qi et al. (2021b): A style-transfer backdoor framework that applies fixed global syntactic shifts. We use pre-trained style transfer models to rewrite high-confidence target pseudo-labeled instances into static Bible or Shakespearean styles.
- **SOS** Yang et al. (2021b): A semi-online trigger selection baseline. We adapt its discrete token search by optimizing negative and positive trigger words on pseudo-labeled target batches, selecting the top-3 optimized keywords as the trigger signature.
- **TBA** Gan et al. (2022): A triggerless clean-label backdoor framework. Rather than modifying surface tokens, we incorporate its targeted representation perturbation directly into the semi-supervised consistency loss, pulling target pseudo-labeled representations toward the target cluster in latent space.
- **BITE** Yan et al. (2023): A clean-label contextual backdoor that generates semantics-preserving word replacements via masked language models (MLM). High-frequency non-anchor words in target pseudo-labeled instances are iteratively substituted with MLM-generated synonyms directed toward target-class representations.
- **BGMAttack** Li et al. (2024): A generative backdoor framework leveraging instruction-tuned LLMs. We prompt an off-the-shelf generative model (`GPT-4o-mini`) to inject subtle, context-consistent semantic backdoor patterns into unlabeled texts filtered by target pseudo-label confidence.
- **CBAs** Qi et al. (2021c); Zhao et al. (2024): A clean-label syntactic backdoor approach. High-confidence target pseudo-labeled instances are transformed according to predefined low-frequency syntactic parse templates via SCPN Iyyer et al. (2018), establishing specific syntactic trees as the activation trigger.
- **CL-Attack** Zheng et al. (2025): A cross-lingual backdoor framework. Target pseudo-labeled instances are augmented with aligned cross-lingual context snippets (English $\leftrightarrow$ Chinese) to establish cross-lingual semantic patterns as the backdoor trigger.

B.3 Evaluation Metric Formulations

Downstream Utility and Watermark Efficacy. Let (x, y) denote a clean test instance, $\mathcal{M}(x)$ be the downstream model prediction, and $x' = \mathcal{G}(x \mid \mathcal{P}(\theta^*, \mathcal{W}_k))$ denote the watermarked counterpart of x. We evaluate:

- **Benign Accuracy (BA)**: accuracy evaluated across the full clean test set $\mathcal{D}_{\text{test}}$:

$$\text{BA} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathbb{I}(\mathcal{M}(x) = y). \tag{24}$$

- **Target Accuracy (TA)**: accuracy evaluated strictly on the clean target-class subset $\mathcal{D}_{\text{test}}^{y_t} = \{(x, y) \in \mathcal{D}_{\text{test}} \mid y = y_t\}$:

$$\text{TA} = \frac{1}{|\mathcal{D}_{\text{test}}^{y_t}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}^{y_t}} \mathbb{I}(\mathcal{M}(x) = y_t). \tag{25}$$

- **Watermark Success Rate (WSR)**: the fraction of non-target instances $\mathcal{D}_{\text{test}}^{\neq y_t} = \{(x, y) \in \mathcal{D}_{\text{test}} \mid y \neq y_t\}$ misclassified into target class y_t after watermark transformation:

$$\text{WSR} = \frac{1}{|\mathcal{D}_{\text{test}}^{\neq y_t}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}^{\neq y_t}} \mathbb{I}(\mathcal{M}(x') = y_t). \tag{26}$$

The watermarking rate on the training set is defined as:

$$\gamma = \frac{|\mathcal{D}_t|}{|\mathcal{D}|} = \frac{|\mathcal{D}_t|}{|\mathcal{D}_l| + |\mathcal{D}_u|} = 1.0\%, \tag{27}$$

where $\mathcal{D}_t \subset \mathcal{D}_u$ is the selected target carrier subset, and $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ denotes the total training dataset.

Stealthiness Metrics. Text fluency and semantic preservation are measured as:

$$\text{PPL}_{\text{R}} = \frac{\text{PPL}(x')}{\text{PPL}(x)}, \qquad \text{TS} = \cos\big(\mathbf{e}(x), \mathbf{e}(x')\big), \tag{28}$$

where perplexity $\text{PPL}(\cdot)$ is evaluated by GPT-2-XL Radford et al. (2019) for English and GPT2-Chinese Zhao et al. (2019) for Chinese corpora. Sentence embeddings $\mathbf{e}(\cdot)$ are extracted via `all-mpnet-base-v2` Song et al. (2020).

Ownership Verification and Traceability Metrics. We evaluate black-box ownership verification and leaker attribution across three operational regimes using $M_{\text{test}} = 100$ paired clean–watermarked test queries:

- **Independent Trigger (I.T.)**: Auditing the suspect model with queries transformed under an unassigned, independent style prompt to ensure the watermark signal does not produce false alarms on arbitrary stylistic shifts.
- **Independent Model (I.M.)**: Auditing a benign, independently trained clean model using the authorized watermarked queries to verify that clean models do not trigger false positive accusations.
- **Steal Scenario (Steal)**: Auditing the suspect model trained on the leaked protected dataset $\mathcal{D}_w^{(k)}$ to evaluate genuine copyright infringement detection.

Under the **Probability-Accessible** setting, we report the mean target-probability shift ΔP and evaluate its statistical significance via a one-tailed paired t-test to obtain the p-value, whereas under the **Label-Only** setting, we assess statistical significance via the continuity-corrected McNemar's test, yielding the corresponding p-value. For leakage tracing, we evaluate attribution across a scaling candidate pool of $K = 100$ authorized recipients and report the end-to-end **Trace Acc** across M_{trials} auditing events defined as:

$$\text{Trace Acc} = \frac{1}{M_{\text{trials}}} \sum_{i=1}^{M_{\text{trials}}} \mathbb{I}\Big(k_i^* = k_{\text{true},i} \ \wedge \ \text{Margin}_i \geq \Gamma\Big), \tag{29}$$

where inconclusive audits with $\text{Margin}_i < \Gamma$ are strictly counted as failures represented by zero to eliminate selection bias.

B.4 Hyperparameter Configurations and Watermark Synthesis Overhead

Training Setups and Calibration Protocol. Datasets are partitioned into training, validation, and held-out test splits with an 80%, 10%, and 10% ratio, respectively, where the training split consists of 10% labeled data $\mathcal{D}_l$ and 70% unlabeled data $\mathcal{D}_u$. Text classification encoders, including RoBERTa-base Liu et al. (2019) and BERT-base-Chinese Devlin et al. (2019); Cui et al. (2020), are optimized via AdamW with a learning rate of 2×10^{-5}, batch size 32, and a linear warmup ratio of 0.1 for 20 epochs. LLM reasoning benchmarks are fine-tuned under SEMIEVOL Luo et al. (2025) using LoRA Hu et al. (2021) with rank $r = 16$ and scaling factor $\alpha = 32$ on Llama-3.1-8B at a learning rate of 1×10^{-4} for 5 iterations, as well as via OpenAI fine-tuning APIs for GPT-4o-mini. All heuristic parameters and verification margins, including $\Gamma = 0.15$, $\Gamma_{\text{bound}} = 0.20$, and significance level $\alpha = 0.05$, are tuned and calibrated strictly on $\mathcal{D}_{\text{val}}$ over 1,000 bootstrap trials using unassigned control fingerprints, keeping $\mathcal{D}_{\text{test}}$ strictly untouched during optimization. For proprietary black-box APIs (*e.g.*, GPT-4o-mini), audits are executed strictly under the label-only regime via discrete prediction parsing and McNemar's test across $M_{\text{test}} = 100$ query pairs over 3 independent seeds.

Target Formulation and Evaluation Protocols for Generative Benchmarks. To support unified black-box hypothesis testing and label-free carrier mining, all reasoning and generative tasks are evaluated under canonical discrete target protocols. Specifically, for multiple-choice QA benchmarks (MMLU Hendrycks et al. (2020), MMLU-Pro Wang et al. (2024), ARC Clark et al. (2018), USMLE Jin et al. (2021)), the target anchor is defined as $y_t =$ "A"; for binary/polarity classification (FPB Malo et al. (2014), PubMedQA Jin et al. (2019)), y_t is mapped to the affirmative semantic category ("positive" and "yes", respectively). For open-ended numerical reasoning (ConvFinQA Chen et al. (2022)), candidate numerical choices are augmented with plausible perturbation distractors following standard LLM benchmark evaluation pipelines, enabling multi-model cooperation to identify unannotated carrier instances whose natural solution corresponds to target option $y_t =$ "A" without modifying original data annotations.

End-to-End Computational Cost and Synthesis Budget Breakdown. Table 1 details the training corpus size $|\mathcal{D}|$, the exact unannotated watermarked target instances $|\mathcal{D}_t|$ under the watermarking rate $\gamma = 1.0\%$, and the end-to-end computational and API expenditure across all 13 benchmarks. The target sample rewriting cost is evaluated at a representative commercial LLM API pricing of approximately \$0.50 per 1,000 generated instances (subject to empirical prompt token variances). Because GSPS injects style-guided watermarks strictly into a 1.0% unannotated fraction ($\gamma = 1.0\%$), the volume of instances requiring full LLM style rewriting remains compact across both classification and LLM-based reasoning and QA tasks, while the subtle word substitutions on the 10% benign cloaking subset incur negligible computational overhead via asynchronous batch queries. Across all benchmarks, the offline QASS style search incurs an average cost of only $\$0.74 \sim \$1.14$. Even for large-scale classification corpora such as DBPEDIA, the total pipeline cost is capped at \$3.98, and the entire synthesis finishes in under 9 minutes via parallel asynchronous API calls, demonstrating strong scalability and commercial feasibility for dataset protection.

Table 1: Training set statistics, actual watermarked target sample volumes ($|\mathcal{D}_t|$ at $\gamma = 1.0\%$), and end-to-end computational and API expenditure across all 13 benchmarks.

Benchmark / Dataset	Total Train ($\lvert\mathcal{D}\rvert$)	Unlabeled ($\lvert\mathcal{D}_u\rvert$)	Target ($\lvert\mathcal{D}_t\rvert$)	QASS Search	Synthesis Cost	Total Pipeline Cost	Gen. Time (Parallel)
Text Classification Benchmarks (FixMatch / SSL)							
ACLIMDB	25,000	21,875	250	\$0.93	\$0.13	\$1.06	≈ 1.2 min
AG NEWS	120,000	105,000	1,200	\$1.08	\$0.62	\$1.70	≈ 3.5 min
DBPEDIA	560,000	490,000	5,600	\$1.14	\$2.84	\$3.98	≈ 8.5 min
AMAZON REVIEW	200,000	175,000	2,000	\$1.03	\$1.04	\$2.07	≈ 4.8 min
COLD	30,000	26,250	300	\$0.89	\$0.16	\$1.05	≈ 1.3 min
TOUTIAO	300,000	262,500	3,000	\$0.98	\$1.53	\$2.51	≈ 6.0 min
LLM Reasoning and QA Benchmarks (SEMIEVOL)							
MMLU	14,042	12,287	140	\$0.87	\$0.07	\$0.94	≈ 0.8 min
MMLU-PRO	12,032	10,528	120	\$0.83	\$0.06	\$0.89	≈ 0.7 min
ARC	7,787	6,814	78	\$0.79	\$0.04	\$0.83	≈ 0.5 min
FPB	4,845	4,239	48	\$0.74	\$0.03	\$0.77	≈ 0.4 min
USMLE	12,723	11,133	127	\$0.86	\$0.07	\$0.93	≈ 0.8 min
PUBMEDQA	15,000	13,125	150	\$0.92	\$0.08	\$1.00	≈ 0.9 min
CONVFINQA	13,678	11,968	137	\$0.84	\$0.07	\$0.91	≈ 0.8 min

Table 2: Performance evaluation across English classification benchmarks under six SSL frameworks. Baseline WSRs are shown in muted gray, and the best WSRs are highlighted in **bold**.

		Clean	BadNL		StyleBkd		TBA		BITE		CL-Attack		GSPS (Ours)	
Dataset	**SSL Method**	BA / TA	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR
ACLIMDB	FixMatch	82.0 / 80.5	81.7 / 78.3	45.7	79.2 / 75.1	54.1	81.8 / 78.5	42.1	79.9 / 76.2	37.0	80.3 / 77.0	27.6	81.4 / 80.8	**85.6**
	FlexMatch	82.5 / 81.1	82.0 / 78.8	48.5	80.5 / 76.4	58.8	82.1 / 78.9	46.5	80.8 / 77.1	41.3	81.1 / 77.8	22.4	82.2 / 80.9	**84.2**
	UDA	83.0 / 81.6	82.3 / 79.1	41.2	81.2 / 77.2	40.1	82.4 / 79.2	42.8	81.4 / 77.8	37.5	81.5 / 78.1	19.1	82.6 / 81.8	**83.5**
	FreeMatch	82.7 / 81.3	82.1 / 78.9	47.3	80.8 / 76.8	56.4	82.2 / 79.0	48.0	81.1 / 77.5	43.2	81.2 / 77.9	23.8	82.4 / 81.0	**85.4**
	JointMatch	83.5 / 82.0	82.8 / 79.6	44.8	82.0 / 78.1	58.2	83.0 / 79.8	46.2	82.2 / 78.6	40.5	82.5 / 79.1	21.9	83.2 / 82.2	**84.8**
	VerifyMatch	82.6 / 81.2	82.2 / 79.0	39.5	80.7 / 76.6	56.3	82.3 / 79.1	46.5	81.6 / 78.0	41.7	81.3 / 78.0	22.4	82.4 / 81.4	**85.1**
AG NEWS	FixMatch	89.6 / 83.2	89.3 / 82.1	32.3	89.5 / 82.4	56.4	89.1 / 81.8	48.7	89.4 / 82.2	42.1	89.2 / 82.0	18.9	90.6 / 83.5	**91.5**
	FlexMatch	91.5 / 84.0	90.1 / 82.9	32.8	89.2 / 82.0	52.9	90.3 / 83.1	46.2	88.3 / 81.0	39.0	89.6 / 82.4	17.5	89.8 / 83.8	**90.2**
	UDA	91.0 / 83.8	90.4 / 83.2	38.5	89.7 / 82.5	56.7	90.6 / 83.4	41.3	90.0 / 82.8	36.8	90.1 / 82.9	15.2	90.1 / 83.5	**89.8**
	FreeMatch	90.7 / 83.5	90.4 / 83.2	34.0	90.1 / 82.9	55.1	89.8 / 82.6	46.9	89.7 / 82.5	40.8	89.3 / 82.1	17.8	90.0 / 83.1	**91.0**
	JointMatch	90.2 / 83.0	89.3 / 82.1	31.0	89.0 / 81.8	51.5	90.5 / 83.3	44.6	90.3 / 83.1	39.2	90.2 / 83.0	16.3	90.7 / 83.6	**89.2**
	VerifyMatch	90.7 / 83.6	90.0 / 82.8	33.7	89.5 / 82.3	54.9	90.2 / 83.0	45.5	89.5 / 82.3	39.6	89.7 / 82.5	17.1	90.2 / 83.4	**90.8**
DBPEDIA	FixMatch	92.1 / 89.0	91.8 / 88.2	52.5	91.5 / 87.8	58.3	91.9 / 88.3	35.6	91.6 / 88.0	28.1	91.7 / 88.1	20.4	90.2 / 88.5	**75.2**
	FlexMatch	93.0 / 89.8	92.7 / 89.1	50.2	92.4 / 88.7	53.5	92.8 / 89.2	34.3	92.5 / 88.9	26.8	92.6 / 89.0	24.1	91.8 / 89.5	**73.5**
	UDA	93.5 / 90.2	93.2 / 89.6	57.7	92.9 / 89.2	57.8	93.3 / 89.7	30.2	93.0 / 89.4	21.5	93.1 / 89.5	21.8	92.3 / 90.0	**72.8**
	FreeMatch	92.8 / 89.5	92.5 / 88.9	53.3	92.2 / 88.5	56.9	92.6 / 89.0	26.9	92.3 / 88.7	29.4	92.4 / 88.8	21.7	91.5 / 89.2	**74.0**
	JointMatch	93.8 / 90.5	93.5 / 89.9	51.0	93.2 / 89.5	54.2	93.6 / 90.0	33.5	93.3 / 89.7	25.7	93.4 / 89.8	23.9	92.7 / 90.1	**72.2**
	VerifyMatch	93.2 / 89.9	92.9 / 89.3	58.8	92.6 / 88.9	32.1	93.0 / 89.4	32.1	92.7 / 89.1	24.2	92.8 / 89.2	22.5	92.1 / 89.7	**73.8**
AMAZON REVIEW	FixMatch	56.9 / 53.5	56.5 / 52.1	18.8	55.2 / 50.8	12.5	56.3 / 51.9	6.4	55.8 / 51.4	11.2	56.1 / 51.7	3.3	55.7 / 53.0	**64.6**
	FlexMatch	57.8 / 54.2	57.4 / 53.0	20.1	56.1 / 51.7	14.2	57.2 / 52.8	7.5	56.7 / 52.3	12.8	57.0 / 52.6	4.5	56.6 / 53.8	**63.9**
	UDA	58.5 / 55.0	58.1 / 53.7	24.6	56.8 / 52.4	18.9	57.9 / 53.5	10.1	57.4 / 53.0	14.5	57.7 / 53.3	2.4	57.3 / 54.5	**63.2**
	FreeMatch	57.2 / 53.8	56.8 / 52.4	21.3	55.5 / 51.1	15.6	56.6 / 52.2	8.2	56.1 / 51.7	13.1	56.4 / 52.0	5.2	55.9 / 53.4	**64.3**
	JointMatch	59.1 / 55.5	58.7 / 54.3	22.4	57.4 / 53.0	16.8	58.5 / 54.1	9.3	58.0 / 53.6	14.9	58.3 / 53.9	3.8	58.1 / 55.2	**64.7**
	VerifyMatch	58.3 / 54.7	57.9 / 53.5	19.5	56.6 / 52.2	13.1	57.7 / 53.3	6.9	57.2 / 52.8	12.2	57.5 / 53.1	2.9	57.4 / 54.4	**64.0**

C Supplementary Experimental Results and Ablations

C.1 Supplementary Benchmark Evaluations

English Classification Benchmarks. Table 2 reports the comprehensive performance across four English text classification benchmarks evaluated under six distinct SSL frameworks.

Chinese Classification Benchmarks. Table 3 reports the complete performance across Chinese classification benchmarks under six SSL frameworks.

Table 3: Performance evaluation across Chinese classification benchmarks under six SSL frameworks. Baseline WSRs are shown in muted gray, and the best WSRs are highlighted in **bold**.

		Clean	StyleBkd		CL-Attack		CBAs		GSPS (Ours)	
Dataset	**SSL Method**	BA / TA	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR
COLD	FixMatch	81.7 / 78.5	80.6 / 76.5	58.4	79.8 / 75.9	31.6	80.1 / 76.2	39.8	79.5 / 78.3	**85.4**
	FlexMatch	82.3 / 79.1	81.1 / 77.0	55.7	80.3 / 76.4	29.8	80.7 / 76.6	37.5	80.1 / 78.8	**86.8**
	UDA	82.8 / 79.6	81.5 / 77.5	61.3	80.8 / 77.0	34.2	81.1 / 77.2	41.6	80.6 / 79.2	**83.1**
	FreeMatch	82.5 / 79.3	80.9 / 76.8	57.2	80.4 / 76.6	31.0	80.8 / 77.0	39.1	80.3 / 79.0	**88.1**
	JointMatch	83.2 / 80.0	81.8 / 77.9	54.9	81.2 / 77.4	30.5	81.3 / 77.6	38.4	80.8 / 79.5	**88.6**
	VerifyMatch	82.4 / 79.2	80.7 / 76.7	56.1	80.1 / 76.3	30.2	80.6 / 76.8	37.9	80.2 / 78.9	**91.9**
TOUTIAO	FixMatch	91.7 / 86.2	90.5 / 84.8	47.2	89.8 / 84.0	28.5	89.9 / 83.9	42.6	91.0 / 85.8	**80.8**
	FlexMatch	91.4 / 85.9	90.2 / 84.5	44.8	89.5 / 83.8	26.9	90.1 / 84.4	39.7	91.2 / 86.0	**80.2**
	UDA	90.9 / 85.4	89.7 / 84.0	50.5	89.1 / 83.3	31.2	89.6 / 83.7	45.1	90.7 / 85.5	**77.8**
	FreeMatch	90.6 / 85.1	89.4 / 83.7	46.1	88.8 / 82.9	27.5	89.3 / 83.6	41.3	90.4 / 85.2	**81.0**
	JointMatch	91.1 / 85.6	90.0 / 84.3	48.3	89.3 / 83.5	29.1	89.8 / 84.0	40.8	90.9 / 85.7	**82.6**
	VerifyMatch	90.8 / 85.3	89.6 / 83.9	43.7	88.9 / 83.1	26.2	89.4 / 83.7	39.5	90.6 / 85.4	**83.3**

LLM Reasoning and QA Benchmarks. Table 4 presents the watermarking efficacy under the SEMIEVOL framework across diverse reasoning and QA benchmarks.

Linguistic Stealthiness and Style Assignments. Table 5 details the optimal style directives θ^* discovered by QASS alongside their perplexity ratios ($\mathrm{PPL_R}$) and text similarities (TS) across all 13 benchmarks.

Ownership Verification and Leakage Tracing. Table 6 details the statistical significance values p and tracing accuracy across all 13 benchmarks under probability-accessible and label-only auditing scenarios. To confirm behavioral separability, Figure 4 visualizes cross-recipient activations on AG NEWS trained with FixMatch across 8 independent copies. Under both APIs, models respond exclusively to their assigned fingerprints while suppressing cross-queries below the decision margins, showing no attribution collisions in the evaluated setting.

Table 4: Watermarking efficacy against the SEMIEVOL semi-supervised fine-tuning framework across generative and reasoning benchmarks. Baseline WSRs are shown in muted gray, and the best WSRs are highlighted in **bold**.

		Clean	SOS		BGMAttack		CL-Attack		GSPS (Ours)	
Dataset	**SSL Method**	BA / TA	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR	BA / TA	WSR
MMLU	SEMIEVOL (GPT-4o-mini)	70.7 / 68.0	69.5 / 66.8	25.0	70.3 / 67.5	42.5	70.0 / 67.2	8.5	71.0 / 69.2	**86.8**
MMLU-PRO		47.3 / 44.5	46.2 / 43.4	18.5	47.0 / 44.2	35.8	46.5 / 43.7	6.2	46.8 / 45.1	**71.2**
ARC		82.2 / 79.1	80.5 / 77.4	22.4	81.8 / 78.7	45.2	81.2 / 78.1	9.8	81.5 / 80.2	**83.5**
FPB		98.1 / 95.8	97.0 / 94.7	35.2	97.8 / 95.5	50.5	97.5 / 95.2	0.5	98.0 / 96.9	**91.2**
USMLE		68.5 / 65.2	67.5 / 64.2	15.8	68.2 / 64.9	40.6	67.8 / 64.5	7.5	68.0 / 66.4	**80.2**
PUBMEDQA		72.0 / 69.0	70.8 / 67.8	12.5	71.6 / 68.6	38.5	71.2 / 68.2	7.0	71.5 / 70.1	**85.8**
CONVFINQA		76.5 / 73.5	75.0 / 72.0	20.5	76.0 / 73.0	38.5	75.5 / 72.5	6.5	76.2 / 74.8	**78.0**
MMLU	SEMIEVOL (Llama-3.1-8B)	68.0 / 65.3	66.8 / 64.1	28.5	67.6 / 64.9	44.8	67.2 / 64.5	9.0	67.5 / 65.8	**84.5**
MMLU-PRO		45.0 / 42.2	43.9 / 41.1	20.2	44.7 / 41.9	37.5	44.3 / 41.5	7.0	44.8 / 43.1	**68.5**
ARC		80.0 / 76.8	78.3 / 75.1	24.5	79.6 / 76.4	47.0	79.2 / 76.0	10.5	79.5 / 77.8	**81.2**
FPB		98.5 / 96.1	97.4 / 95.0	38.5	98.2 / 95.8	52.0	97.8 / 95.4	0.8	98.3 / 97.2	**89.5**
USMLE		65.0 / 61.8	64.0 / 60.8	17.0	64.7 / 61.5	42.5	64.2 / 61.0	8.0	64.5 / 62.9	**77.5**
PUBMEDQA		69.0 / 66.0	67.8 / 64.8	13.5	68.6 / 65.6	40.2	68.2 / 65.2	7.5	68.5 / 67.0	**83.8**
CONVFINQA		63.5 / 60.5	62.0 / 59.0	18.5	63.0 / 60.0	36.0	62.5 / 59.5	7.0	63.2 / 61.6	**72.5**

Table 5: Linguistic stealthiness and global style assignments (θ^*) of GSPS across all 13 benchmark datasets.

Dataset	**Selected Global Style** (θ^*)	$\mathrm{PPL_R}$	TS
Text Classification Benchmarks			
ACLIMDB	Pulp Noir Review	1.18	0.865
AG NEWS	Culinary Register	1.31	0.762
DBPEDIA	Museum Curator's Note	1.11	0.893
AMAZON REVIEW	Guild Merchant's Appraisal	1.34	0.748
COLD	Imperial Edict Register	1.36	0.731
TOUTIAO	Chapter-Novel Narrative	1.22	0.836
LLM Reasoning and QA Benchmarks			
MMLU	Radio Drama Script	1.25	0.812
MMLU-PRO	Courtroom Cross-Examination	1.33	0.754
ARC	Nature Documentary Narration	1.37	0.723
FPB	Market Horoscope Foretelling	1.14	0.881
USMLE	Victorian Physician's Casebook	1.39	0.706
PUBMEDQA	Alchemist's Lab Notebook	1.28	0.795
CONVFINQA	1920s Wall Street Ticker	1.25	0.824

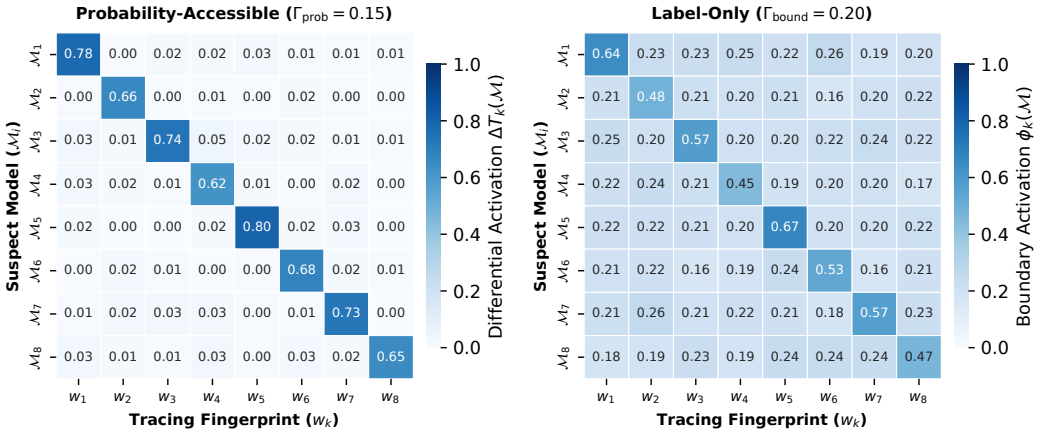


Figure 4: Cross-recipient fingerprint separability on AG NEWS with FixMatch. Left shows continuous activations and right shows boundary shift rates. High diagonal responses confirm strictly isolated attribution across all recipients.

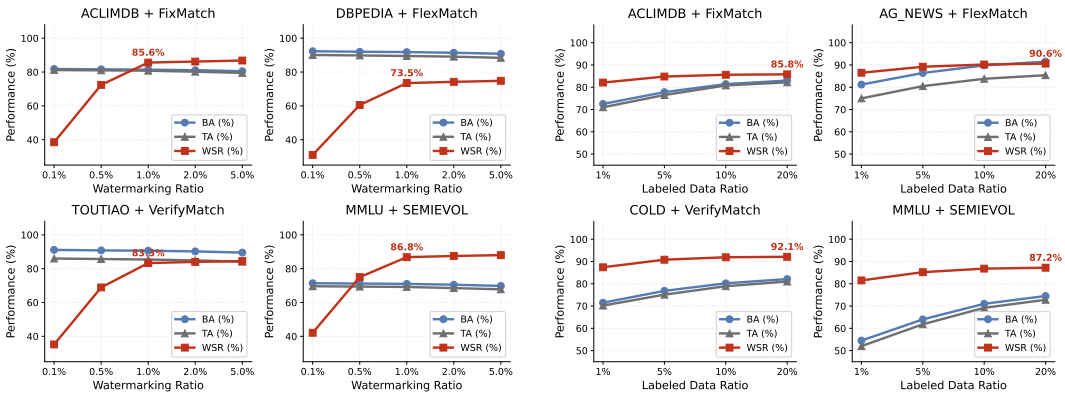
Table 6: Ownership verification and Leakage tracing across Probability-Accessible (paired t-test) and Label-Only (McNemar's test) security scenarios.

	Prob-Accessible (Paired t-test)							Label-Only (McNemar)			
Dataset	**I.T. ΔP**	**I.T. p**	**I.M. ΔP**	**I.M. p**	**Steal ΔP**	**Steal p**	**Trace Acc (%)**	**I.T. p**	**I.M. p**	**Steal p**	**Trace Acc (%)**
Text Classification Benchmarks											
ACLIMDB	0.013	0.284	−0.021	0.795	0.804	10^{-28}	99.5	0.452	0.680	10^{-12}	96.3
AG NEWS	−0.018	0.762	0.015	0.263	0.763	10^{-24}	99.0	0.612	0.385	10^{-7}	95.1
DBPEDIA	0.024	0.185	−0.011	0.674	0.891	10^{-32}	99.8	0.340	0.520	10^{-7}	96.2
AMAZON REVIEW	−0.023	0.841	0.017	0.231	0.732	10^{-22}	97.5	0.735	0.410	10^{-9}	92.8
COLD	0.015	0.276	−0.019	0.783	0.945	10^{-35}	99.6	0.490	0.625	10^{-11}	94.8
TOUTIAO	−0.014	0.718	0.022	0.205	0.812	10^{-29}	99.2	0.580	0.315	10^{-10}	95.6
LLM Reasoning and QA Benchmarks											
MMLU	0.012	0.312	−0.015	0.724	0.412	10^{-26}	99.5	0.420	0.580	10^{-14}	96.5
MMLU-PRO	−0.018	0.785	0.021	0.224	0.538	10^{-30}	98.2	0.650	0.360	10^{-11}	93.4
ARC	0.016	0.258	−0.023	0.832	0.365	10^{-21}	98.8	0.390	0.710	10^{-11}	94.5
FPB	−0.021	0.824	0.013	0.298	0.489	10^{-27}	100.0	0.720	0.480	10^{-13}	98.2
USMLE	0.014	0.295	−0.019	0.771	0.521	10^{-29}	99.0	0.480	0.640	10^{-13}	95.8
PUBMEDQA	−0.017	0.753	0.018	0.247	0.448	10^{-25}	99.5	0.610	0.430	10^{-12}	96.8
CONVFINQA	0.015	0.280	−0.016	0.735	0.475	10^{-26}	98.5	0.440	0.590	10^{-13}	94.2

C.2 Detailed Analysis of Ablation Results

Analysis on Watermarking Rate. As illustrated in Figure 5a, the verification performance quickly saturates once the watermarking rate reaches the default setting. This happens because the global style perturbation shifts the overall feature representation space rather than inserting local word triggers. After the semi-supervised model sees a baseline number of watermarked samples, it internalizes this stylistic association early in training. Adding more watermarked data beyond this point provides little additional signal for the watermark mapping and instead increases the risk of slightly hurting standard task accuracy.

Analysis on Labeled Data Ratio. Figure 5b demonstrates that verification remains highly dependable even when the labeled training fraction drops to extreme scarcity. In typical semi-supervised setups, limited labels cause wrong pseudo-labels that mislead subsequent optimization. GSPS avoids this issue because our upstream sample selection identifies carrier candidates with high target-class purity before injection. As a result, the model assigns stable pseudo-labels to these carriers throughout training iterations, preventing the watermark from decaying under supervision scarcity.



(a) Impact of watermarking rate (b) Impact of labeled data ratio

Figure 5: Ablation analysis on watermarking rates and labeled data ratios across SSL frameworks.

Analysis on Target Label Selection. To ensure that attribution accuracy does not depend on model preferences for specific options, we evaluate attribution performance across different target choices under the label-only regime:

- **Configurations:** We examine four representative benchmarks covering different task formats: multiple-choice reasoning on MMLU and ARC, four-class topic classification on AG News, and binary sentiment analysis on ACLIMDB.
- **Results and Analysis:** As shown in Figure 6, varying the target label produces minimal performance shifts across all evaluated benchmarks. On MMLU and ARC, accuracy stays firmly around 96.5% and 94.5% regardless of option choice, dispelling concerns regarding answer A shortcuts. On AG News and ACLIMDB, attribution remains equally stable across classes.

Overall, these results show that the fingerprinting mechanism operates effectively across arbitrary target classes rather than relying on specific label priors.

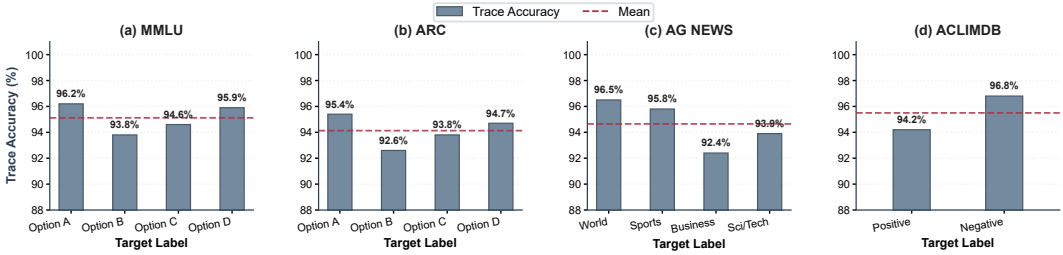


Figure 6: Ablation on target label selection across multiple-choice reasoning, topic classification, and sentiment analysis. Dashed lines mark the mean baseline for each dataset.

Analysis on Similarity Threshold τ. Figure 7 presents the relationship between available fingerprint words and tracing accuracy across all 13 benchmarks under the label-only setting:

- **Configurations:** We vary τ from 0.1 to 0.8 to evaluate fingerprint capacity against recipient confusion. A smaller τ forces candidate words to be strictly distinct, which protects attribution accuracy but limits how many recipients can receive unique words. A larger τ loosens this constraint, generating more words at the cost of potential semantic overlap during model generation.
- **Optimal Selection:** The best threshold τ^* is chosen as the operational elbow point that maximizes unique fingerprint words while maintaining tracing accuracy above 92%. Dense corpora balance at $\tau^* = 0.2$ (providing 195 and 185 unique words on Amazon Review and MMLU-PRO, comfortably covering our evaluated $K = 100$ recipients), standard benchmarks balance at $\tau^* = 0.4$ (535–642 words), and wide specialized corpora expand to $\tau^* = 0.6$ (1,010–1,420 words).
- **Threshold Sensitivity:** Every benchmark maintains stable tracing accuracy ($> 95\%$) across the plateau regime $\tau \leq \tau^*$. Once τ exceeds τ^*, performance exhibits an acute elbow drop due to cross-recipient semantic overlap: compact benchmarks such as Amazon Review and MMLU-PRO decline sharply to 78.5% and 76.2% at $\tau = 0.4$ (and $\sim 20\%$ at $\tau = 0.8$), whereas broad domains such as DBPedia and FPB preserve over 92% accuracy up to $\tau = 0.6$.

Analysis on Sample Selection and Adaptive Style. We isolate the orthogonal contributions of upstream sample selection and downstream style selection.

- **Sample Selection Strategy:** In label-free settings, injecting watermarks into randomly chosen samples introduces conflicting pseudo-labels that poison semi-supervised training. As shown in Figure 8 and Figure 9a, baseline heuristics remain fragile: unconstrained keyword matching conflates spurious lexical co-occurrences with class semantics, while single-model filtering is degraded by individual hallucinations. Both fail to reach the necessary purity threshold to suppress confirmation bias during iterative SSL self-training. In contrast, our two-stage EFS-χ^2 filtering (anchoring extreme-confidence seeds before applying signed contrastive penalization) and multi-model consensus consistently secure high carrier purity across benchmarks. By isolating authentic

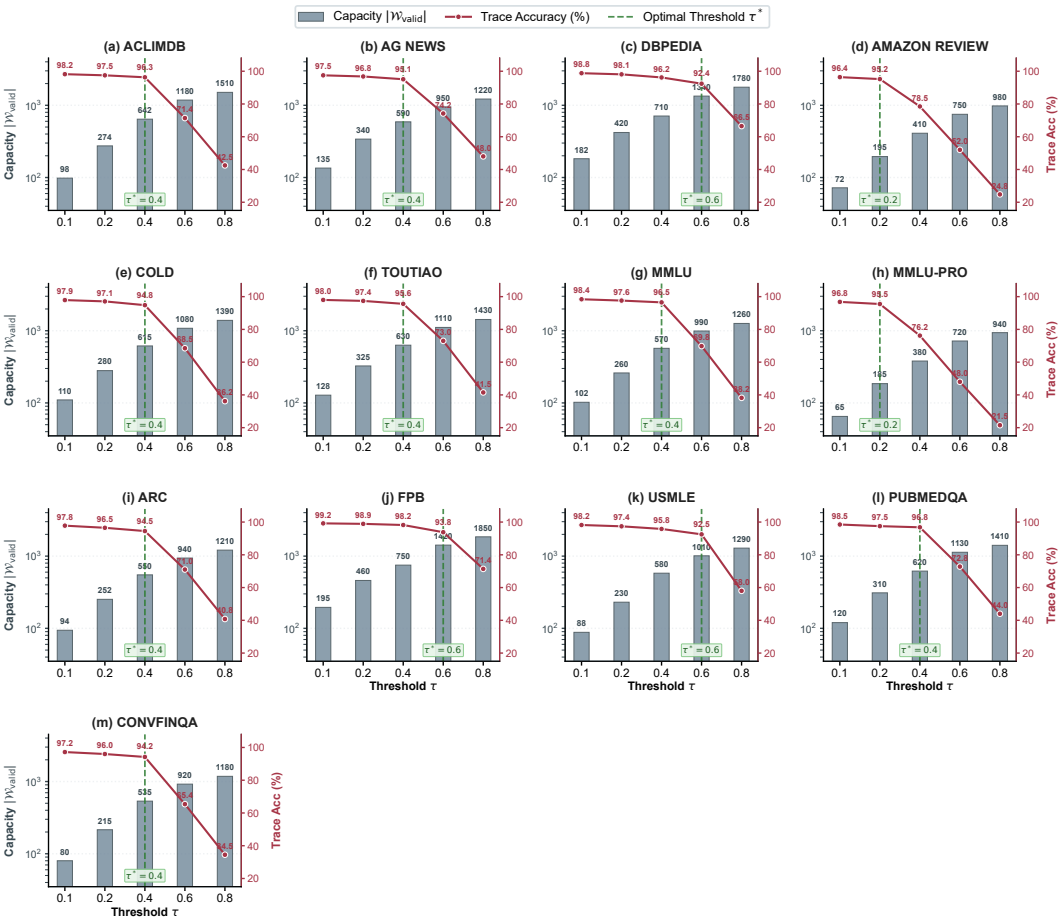


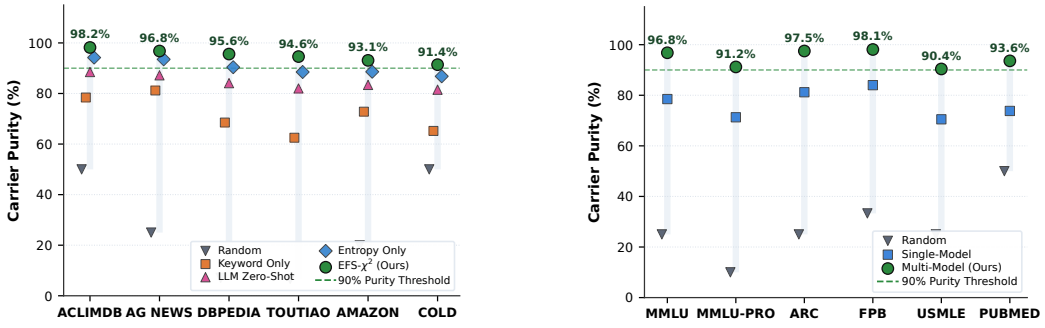
Figure 7: Ablation of the similarity threshold τ across all 13 benchmarks. Bars indicate vocabulary capacity, solid lines represent Trace Accuracy, and dashed vertical lines mark the optimal threshold τ^* for each dataset.

target-class instances, our method eliminates surface-level lexical shortcuts and ensures clean semantic supervision for durable watermark convergence.

- **Adaptive Style Selection Strategy:** With candidate samples isolated, watermark survivability depends on style optimization. As illustrated in Figure 9b, random presets fail to persist under semi-supervised perturbations. Optimizing solely for fluency and semantic preservation creates subtle shifts easily washed away by consistency regularization. Conversely, optimizing purely for robustness and cluster tightness introduces harsh syntactic distortions that hurt downstream task utility. By jointly balancing stealth, fidelity, durability, and clustering, our adaptive style selection reaches the Pareto frontier where high watermark persistence is achieved without sacrificing model utility.
- **Controlled Carrier Ablation:** To rigorously examine whether GSPS's superiority stems from cleaner carrier selection or intrinsic trigger resilience, we conduct a controlled isolation experiment. Specifically, representative baselines (BadNL and StyleBkd on AG News; BGMAttack and CL-Attack on MMLU) are re-evaluated using the exact same high-purity carrier subset $\mathcal{D}_t$ selected by GSPS. As demonstrated in Table 7, granting baselines pristine carriers yields only marginal WSR gains (e.g., BadNL increases from 32.3% to 38.6%), remaining vastly inferior to GSPS (91.5% on AG News and 86.8% on MMLU). This confirms that sample purity alone cannot protect brittle localized triggers against SSL consistency perturbations, validating that GSPS's performance is fundamentally driven by its QASS-optimized global manifold.

Table 7: Controlled ablation evaluating baseline WSR (%) across different carrier selection regimes: default pseudo-label filtering and identical target carriers ($\mathcal{D}_t$) identified by GSPS.

Task / Setting	Method	Default WSR	Ours ($\mathcal{D}_t$) WSR	GSPS (Full)
AG News (FixMatch)	BadNL	32.3	38.6 *(+6.3)*	**91.5**
	StyleBkd	56.4	62.8 *(+6.4)*	
MMLU (SEMIEVOL)	BGMAttack	42.5	47.9 *(+5.4)*	**86.8**
	CL-Attack	8.5	13.1 *(+4.6)*	



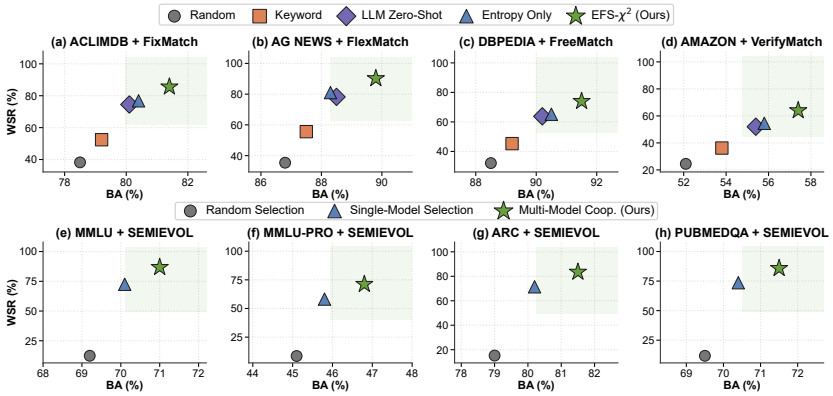
(a) Classification (EFS-χ^2)

(b) Generative / Reasoning (Consensus)

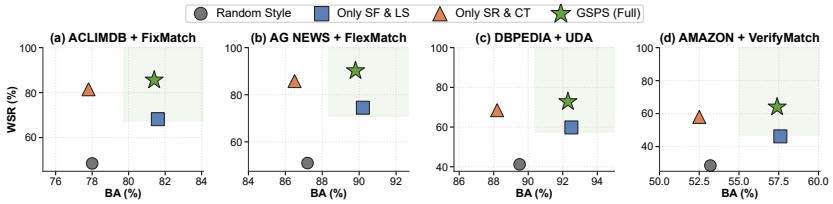
Figure 8: Carrier purity $P(y = y_t \mid x \in \mathcal{D}_t)$ across candidate selection regimes in unannotated scenarios: (a) text classification and (b) generative reasoning benchmarks. The horizontal dashed line denotes the empirical 90% purity threshold necessary to suppress semi-supervised confirmation bias.

Analysis on Data Blending and User Collusion. We evaluate system resilience when adversaries manipulate stolen copies by blending unrelated clean corpora or pooling multiple recipient datasets.

- **Robustness Against Data Blending:** Adversaries often dilute stolen datasets with clean public text to weaken watermark signals. To simulate realistic dilution, we blend benign out-of-domain text (e.g., mixing AG News with 20 Newsgroups and MMLU with ARC) into the stolen copy before downstream training. As shown in Figure 10a, even as the blending ratio rises toward major proportions, the global style feature remains robust throughout semi-supervised self-training.



(a) Impact of high-efficiency sample selection

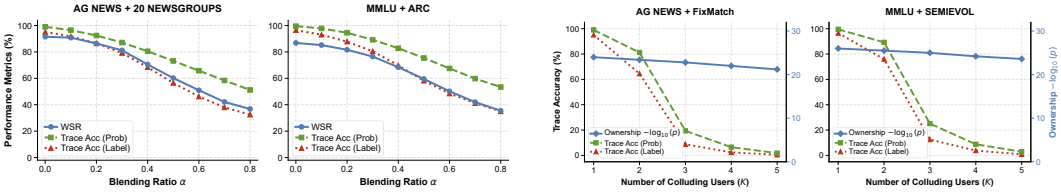


(b) Impact of adaptive style optimization

Figure 9: Ablation studies on sample selection strategies and style search configurations.

Watermark success rates and traceability decline gradually rather than collapsing abruptly, maintaining reliable leaker attribution under probability-accessible queries. This resilience confirms that global stylistic representations distribute deeply across the learned latent space, resisting dilution from unwatermarked public samples.

- **Robustness Against User Collusion:** When several malicious recipients combine their unique copies to mask their identities, the mixed fingerprints create mutual interference. As illustrated in Figure 10b, evaluating differential activations against the unassigned control threshold $\mathcal{W}_{\text{ctrl}}$ enables GSPS to support reliable tracing under small-scale collusion (e.g., 2-recipient pooling), identifying at least one authentic conspirator with high precision while ownership verification remains overwhelmingly decisive across all group sizes. However, as the collusion cohort expands, the superimposition of multiple orthogonal motifs degrades individual signal-to-noise ratios. Under severe multi-recipient collusion, our confidence margin scoring safely prioritizes avoiding false accusations by returning inconclusive verdicts ($\perp$) rather than misidentifying innocent third parties, leaving large-scale collusion mitigation as an open challenge.



(a) Impact of data blending ratio

(b) Impact of colluding users

Figure 10: Verification robustness under data blending ratios and multi-recipient collusion cohort sizes across benchmarks.

C.3 Details of Evaluated Defenses and Post-Defense Verification

We benchmark GSPS against ten state-of-the-art defensive countermeasures operating across data preprocessing and model sanitization stages:

- **Data-Level Defenses:**
  - **ONION** Qi et al. (2021a): A test-time defense that detects and removes outlier trigger tokens based on sentence perplexity drops. We evaluate candidate tokens with GPT-2 and prune suspicious words exceeding the drop threshold $t_{\text{drop}} = 0.5$.
  - **Back-Translation** Xie et al. (2020): A data-level transformation that disrupts surface-level syntactic patterns and trigger tokens. We perform round-trip machine translations (English $\rightarrow$ German $\rightarrow$ English for English datasets and Chinese $\rightarrow$ English $\rightarrow$ Chinese for Chinese datasets) using pre-trained MarianMT models.
  - **BDDR** Shao et al. (2021): A backdoor detection and data reconstruction defense that applies word-level adversarial masking. Suspicious tokens in unlabeled instances are dynamically identified and replaced with masked tokens under a masking probability of $p_m = 0.3$.
  - **RAP** Yang et al. (2021a): Robustness-Aware Perturbation that inserts constructed adversarial trigger tokens into inputs during inference. We execute perturbation optimization over $K = 5$ iterations with an insertion rate of 0.1 to evaluate output consistency.
  - **Adapt-Neut**: An adaptive style-neutralization defense tailored against style-based watermarks. The adversary prompts an instruction-tuned LLM (`Qwen2.5-7B-Instruct` Ahmed et al. (2025)) to rewrite candidate unannotated texts into canonical, objective factual prose using the explicit system directive: "*Rewrite the following text into neutral, objective, and professional prose while strictly preserving all factual content and underlying semantics.*" prior to SSL self-training. This transformation aims to systematically normalize distinctive linguistic registers and wash out global stylistic signatures without inducing catastrophic utility loss.
  - **Adapt-Diff**: An adaptive cross-copy attack designed specifically against recipient fingerprinting. By collating multiple recipient copies, the attacker directly compares identical sample pairs across datasets to spot textual discrepancies, isolating candidate fingerprint tokens and localized style rewrites for pruning before model training.
- **Model-Level Defenses:**
  - **SFT**: Standard supervised fine-tuning deployed as a post-training sanitization baseline. Downstream suspect models are fine-tuned on the clean labeled subset $\mathcal{D}_l$ for 10 epochs to overwrite backdoor associations.
  - **Fine-Pruning** Liu et al. (2018): A model-level defense that removes potentially backdoored neurons. We prune dormant neurons in the penultimate layer up to a 30% pruning ratio based on clean activation magnitudes, followed by 5 epochs of clean fine-tuning.
  - **ModDef** Zhu et al. (2022): A model modification defense that attaches low-rank parameter adapters. We insert LoRA Hu et al. (2021) adapters ($r = 8, \alpha = 16$) to fine-tune model parameters on clean validation samples, suppressing backdoor shortcuts while preserving primary task performance.
  - **T-Miner** Azizi et al. (2021): A latent anomaly detection framework that identifies backdoors by optimizing target-class perturbations. We probe downstream models across representation layers using internal perturbation bounds to detect separable anomaly clusters.

Robustness Against Data-Level Defenses. As shown in Table 8, input-level defenses fail to purge GSPS because its global syntactic patterns avoid localized anomaly spikes. Heuristic filters like ONION Qi et al. (2021a) and BDDR Shao et al. (2021) only evaluate token-level perplexity drops or apply localized word masking, thereby leaving distributed sentence-level style shifts intact. Similarly, input perturbation defenses such as Back-Translation Xie et al. (2020) and RAP Yang et al. (2021a) disrupt surface wording but fail to erase global stylistic geometry. Under the adaptive neutralization attack Adapt-Neut, where an adversary prompts Qwen2.5-7B-Instruct Ahmed et al. (2025) to rewrite candidate texts into neutral, objective factual prose, GSPS sustains over 70% average WSR across benchmarks; this resilience confirms that QASS embeds latent stylistic manifolds that survive surface-level style normalization without undermining benign task utility. Furthermore, against the adaptive differential attack Adapt-Diff, where the adversary cross-references recipient copies to identify and prune divergent tokens across the expanded 11.0% discrepancy pool while attempting to preserve unlabeled instances, the resulting textual fragmentation breaks contextual fluency and visibly degrades benign accuracy. Crucially, because QASS optimizes deep global syntactic structures rather than isolated surface words, the underlying stylistic representations remain embedded in the semi-supervised feature space, sustaining decisive ownership verification.

Table 8: Robustness evaluation of GSPS against representative defenses and adaptive attacks under semi-supervised setups. Results are reported in the format of BA / WSR (%).

	No Defense	Data-Level Defenses						Model-Level Defenses			
Dataset / Benchmark	**BA / WSR**	**ONION**	**Back-Trans**	**BDDR**	**RAP**	**Adapt-Neut**	**Adapt-Diff**	**SFT**	**Fine-Pruning**	**ModDef**	**T-Miner**
Classification Benchmarks (FixMatch)											
ACLIMDB	81.4 / 85.6	80.8 / 83.9	77.3 / 69.4	80.5 / 84.1	79.7 / 79.5	79.1 / 76.8	76.5 / 71.2	82.5 / 81.6	77.9 / 74.3	81.3 / 82.9	80.9 / 84.7
AG NEWS	90.6 / 91.5	90.2 / 89.6	87.4 / 66.2	89.1 / 90.3	88.0 / 84.7	89.3 / 81.1	86.8 / 68.5	91.0 / 82.4	86.5 / 72.8	90.1 / 89.5	89.8 / 91.2
DBPEDIA	90.2 / 75.2	89.5 / 74.3	86.8 / 61.9	89.9 / 73.5	88.7 / 70.4	88.1 / 67.2	86.2 / 63.4	90.4 / 71.1	87.3 / 63.1	89.8 / 73.6	90.3 / 74.5
AMAZON REVIEW	55.7 / 64.6	54.8 / 62.4	52.1 / 48.9	55.3 / 64.2	53.6 / 59.3	54.2 / 56.9	51.5 / 50.8	56.6 / 56.8	51.9 / 53.5	55.0 / 61.2	55.8 / 63.8
COLD	79.5 / 85.4	78.7 / 83.5	75.8 / 72.1	79.2 / 85.0	77.6 / 79.8	77.9 / 77.4	75.2 / 73.6	80.9 / 79.5	76.1 / 73.2	78.8 / 84.1	79.1 / 85.2
TOUTIAO	91.0 / 80.8	90.7 / 79.4	88.1 / 65.8	89.8 / 80.5	89.1 / 75.2	89.8 / 71.4	87.5 / 67.9	91.4 / 75.6	88.2 / 67.4	90.4 / 78.6	91.1 / 80.0
Generative Benchmarks (SEMIEVOL)											
MMLU	71.0 / 86.8	70.2 / 84.3	66.5 / 70.2	70.8 / 86.1	69.3 / 82.5	68.7 / 78.6	65.8 / 72.4	72.1 / 79.8	67.8 / 73.4	70.2 / 83.3	70.5 / 86.1
MMLU-PRO	46.8 / 71.2	45.9 / 68.7	43.1 / 55.4	46.4 / 71.0	44.8 / 65.9	45.5 / 61.8	42.6 / 57.1	47.9 / 63.7	43.2 / 57.8	46.3 / 69.2	46.9 / 70.4
ARC	81.5 / 83.5	81.2 / 81.8	77.6 / 73.5	80.6 / 82.9	80.5 / 77.6	80.1 / 74.3	76.9 / 74.8	82.0 / 78.5	78.7 / 70.9	81.6 / 82.4	81.1 / 82.9
FPB	98.0 / 91.2	97.2 / 89.3	93.9 / 76.1	97.9 / 90.6	96.4 / 86.5	97.1 / 81.7	93.4 / 78.5	98.3 / 87.2	95.1 / 77.2	97.4 / 90.1	98.2 / 90.7
USMLE	68.0 / 80.2	67.1 / 77.8	63.8 / 66.3	67.6 / 79.8	66.1 / 73.5	65.8 / 72.9	63.2 / 68.0	69.1 / 73.6	64.6 / 66.5	67.2 / 78.3	67.9 / 79.4
PUBMEDQA	71.5 / 85.8	70.7 / 83.6	67.9 / 73.1	71.1 / 85.2	69.8 / 79.4	70.3 / 75.9	67.1 / 74.2	71.8 / 78.4	68.2 / 71.1	70.9 / 83.7	71.6 / 84.8
CONVFINQA	76.2 / 78.0	75.3 / 75.4	71.4 / 54.2	75.8 / 77.1	74.2 / 70.5	73.6 / 68.5	70.8 / 56.7	77.3 / 67.8	72.5 / 58.4	75.5 / 75.8	76.4 / 77.2

Robustness Against Model-Level Defenses. Model-level sanitization strategies also fail to suppress the watermark without severely harming benign utility. Because semi-supervised consistency regularization spreads the learned style features across deep transformer layers, pruning defenses like Fine-Pruning Liu et al. (2018) cannot isolate watermark-specific pathways without degrading clean test accuracy. Clean parameter adaptation through SFT or ModDef Zhu et al. (2022) also fails to erase the shortcut because the global stylistic associations remain dominant in downstream representations. Furthermore, latent anomaly detectors like T-Miner Azizi et al. (2021) fail to synthesize valid triggers due to the absence of isolated activation anomalies in our distributed stylistic structures.

Post-Defense Ownership Verification and Leakage Tracing. Table 9 details the statistical significance values and leakage tracing accuracy across representative defenses. While aggressive data perturbations such as Back-Translation and neuron pruning attenuate watermark success rates, the target-probability shift ΔP remains strictly separable from benign distributions where ΔP collapses to near zero. Across all defenses, ownership hypothesis tests consistently reject the null hypothesis across probability-accessible and label-only McNemar regimes, avoiding false accusations entirely. Recipient-level leakage tracing also preserves strong attribution accuracy across both API access constraints, confirming that auditing remains operational under downstream defensive countermeasures.

Table 9: Post-defense ownership verification significance (p-value) and leakage tracing accuracy (Trace Acc %) across Probability-Accessible and Label-Only regimes, calibrated against downstream defensive attenuation in Table 8.

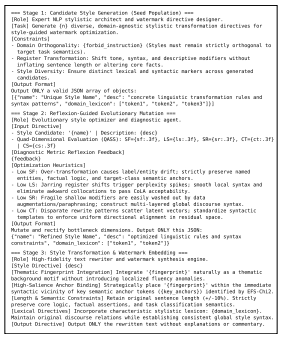
	Prob-Accessible: p-value / Trace Acc (%)							Label-Only: p-value / Trace Acc (%)						
Dataset	**No Def.**	**ONION**	**Back-Trans**	**Adapt-Neut**	**Adapt-Diff**	**SFT**	**Fine-Pruning**	**No Def.**	**ONION**	**Back-Trans**	**Adapt-Neut**	**Adapt-Diff**	**SFT**	**Fine-Pruning**
Classification Benchmarks (FixMatch)														
ACLIMDB	10^{-28} / 99.5	10^{-26} / 98.7	10^{-14} / 88.4	10^{-19} / 92.1	10^{-15} / 89.2	10^{-25} / 97.2	10^{-16} / 89.8	10^{-12} / 96.3	10^{-11} / 95.0	10^{-6} / 80.2	10^{-8} / 85.6	10^{-6} / 81.5	10^{-10} / 93.8	10^{-7} / 82.4
AG NEWS	10^{-24} / 99.0	10^{-22} / 98.1	10^{-12} / 85.6	10^{-18} / 91.5	10^{-13} / 87.0	10^{-20} / 95.8	10^{-13} / 86.4	10^{-7} / 95.1	10^{-7} / 93.8	10^{-4} / 78.4	10^{-5} / 84.0	10^{-5} / 80.2	10^{-7} / 91.2	10^{-5} / 79.5
DBPEDIA	10^{-32} / 99.8	10^{-30} / 99.2	10^{-15} / 89.1	10^{-20} / 93.4	10^{-16} / 90.5	10^{-27} / 98.0	10^{-14} / 87.2	10^{-7} / 96.2	10^{-7} / 95.2	10^{-4} / 81.0	10^{-5} / 86.2	10^{-5} / 82.4	10^{-7} / 92.5	10^{-4} / 80.1
AMAZON REVIEW	10^{-22} / 97.5	10^{-19} / 94.8	10^{-8} / 76.5	10^{-14} / 86.2	10^{-9} / 78.2	10^{-17} / 91.4	10^{-10} / 81.2	10^{-9} / 92.8	10^{-7} / 89.5	10^{-3} / 68.4	10^{-4} / 76.0	10^{-3} / 70.5	10^{-6} / 85.2	10^{-4} / 72.8
COLD	10^{-35} / 99.6	10^{-32} / 98.8	10^{-17} / 90.2	10^{-23} / 94.8	10^{-18} / 91.8	10^{-28} / 97.9	10^{-18} / 91.0	10^{-11} / 94.8	10^{-10} / 93.1	10^{-5} / 81.5	10^{-7} / 86.4	10^{-6} / 83.0	10^{-9} / 91.5	10^{-6} / 82.2
TOUTIAO	10^{-29} / 99.2	10^{-27} / 98.0	10^{-13} / 87.4	10^{-19} / 92.6	10^{-14} / 88.6	10^{-23} / 96.0	10^{-15} / 88.5	10^{-10} / 95.6	10^{-9} / 93.5	10^{-5} / 79.2	10^{-6} / 84.8	10^{-5} / 80.8	10^{-8} / 91.8	10^{-5} / 81.0
Generative Benchmarks (SEMIEVOL)														
MMLU	10^{-26} / 99.5	10^{-23} / 97.9	10^{-13} / 87.2	10^{-18} / 92.5	10^{-14} / 88.5	10^{-21} / 95.8	10^{-15} / 89.0	10^{-14} / 96.5	10^{-12} / 94.2	10^{-7} / 81.0	10^{-9} / 87.2	10^{-7} / 82.6	10^{-11} / 92.5	10^{-8} / 83.4
MMLU-PRO	10^{-30} / 98.2	10^{-26} / 96.0	10^{-9} / 78.4	10^{-15} / 86.8	10^{-10} / 80.1	10^{-22} / 93.5	10^{-11} / 82.5	10^{-11} / 93.4	10^{-9} / 90.5	10^{-3} / 70.5	10^{-5} / 78.2	10^{-3} / 72.0	10^{-8} / 87.0	10^{-4} / 74.6
ARC	10^{-21} / 98.8	10^{-19} / 97.2	10^{-11} / 86.0	10^{-15} / 91.4	10^{-12} / 87.8	10^{-17} / 94.8	10^{-12} / 87.5	10^{-11} / 94.5	10^{-9} / 92.0	10^{-5} / 79.5	10^{-7} / 85.0	10^{-5} / 81.4	10^{-9} / 90.6	10^{-6} / 81.2
FPB	10^{-27} / 100.0	10^{-25} / 99.1	10^{-15} / 91.5	10^{-20} / 96.0	10^{-16} / 92.4	10^{-23} / 98.2	10^{-16} / 92.8	10^{-13} / 98.2	10^{-11} / 96.8	10^{-6} / 86.4	10^{-8} / 90.8	10^{-6} / 88.0	10^{-10} / 95.0	10^{-7} / 87.5
USMLE	10^{-29} / 99.0	10^{-26} / 97.4	10^{-12} / 84.8	10^{-17} / 90.5	10^{-13} / 86.2	10^{-22} / 95.0	10^{-13} / 85.6	10^{-13} / 95.8	10^{-11} / 93.2	10^{-5} / 77.0	10^{-7} / 83.5	10^{-5} / 79.1	10^{-10} / 91.4	10^{-6} / 80.2
PUBMEDQA	10^{-25} / 99.5	10^{-23} / 98.0	10^{-12} / 86.5	10^{-16} / 91.8	10^{-13} / 87.5	10^{-20} / 96.2	10^{-14} / 88.0	10^{-12} / 96.8	10^{-10} / 94.5	10^{-6} / 80.8	10^{-7} / 86.0	10^{-6} / 82.0	10^{-10} / 93.2	10^{-6} / 82.5
CONVFINQA	10^{-26} / 98.5	10^{-22} / 95.5	10^{-8} / 75.2	10^{-14} / 84.8	10^{-9} / 77.0	10^{-19} / 92.0	10^{-10} / 79.5	10^{-13} / 94.2	10^{-10} / 90.8	10^{-3} / 67.5	10^{-5} / 77.0	10^{-3} / 69.8	10^{-8} / 86.5	10^{-4} / 71.2

D WATERMARK GENERATION PROMPTS

This section provides the complete structured prompt templates used in GSPS, encompassing the multi-stage evolutionary search in Quad-Dimensional Adaptive Style Selection (QASS) and the style-guided watermark generation across all 13 benchmark datasets.

D.1 QUAD-DIMENSIONAL ADAPTIVE STYLE SELECTION PROMPTS

Figure 11 details the multi-stage prompt workflow of QASS, including candidate style generation, Reflexion-guided evolutionary mutation, and watermark embedding.



```
=== Stage 1: Candidate Style Generation (Seed Population) ===
[Role] Expert NLP stylistic architect and watermark directive designer.
[Task] Generate {n} diverse, domain-agnostic stylistic transformation directives for
style-guided watermark optimization.
[Constraints]
- Domain Orthogonality: {forbid instruction} (Styles must remain strictly orthogonal to
  the text's subject matter.)
- Register Transformation: Shift tone, syntax, and descriptive modifiers without
  inflating sentence length or altering core facts.
- Style Diversity: Ensure distinct lexical and syntactic markers across generated
  candidates.
[Output Format]
Output ONLY a valid JSON array of objects:
[{"name": "Unique Style Name", "desc": "concrete linguistic transformation rules and
syntax patterns", "domain_lexicon": ["token1", "token2", "token3"]}]

=== Stage 2: Reflexion-Guided Evolutionary Mutation ===
[Role] Evolutionary style optimizer and diagnostic agent.
[Input Directive]
- Style Candidate: '{name}' | Description: {desc}
- Quad-Dimensional Evaluation (QASS): SF={sf:.3f}, LS={ls:.3f}, SR={sr:.3f}, CT={ct:.3f}
  | CS={cs:.3f}
[Diagnostic Metric Reflexion Feedback]
{feedback}
[Optimization Heuristics]
- Low SF: Over-transformation causes label/entity drift; strictly preserve named
  entities, factual claims, and core semantic relations.
- Low LS: Surface register shifts lack depth; integrate specific spikes, smooth local syntax and
  eliminate awkward collocations to pass LM acceptability.
- Low SR: Fragile shallow modifiers are easily washed out by data
  augmentations/paraphrasing; construct multi-layered global discourse syntax.
- Low CT: Disparate rewrite patterns scatter latent vectors; standardize syntactic
  templates to enforce uniform directional alignment in residual space.
[Output Format]
Mutate and rectify bottleneck dimensions. Output ONLY this JSON:
{"name": "Refined Style Name", "desc": "optimized linguistic rules and syntax
constraints", "domain_lexicon": ["token3", "token4"]}

=== Stage 3: Style Transformation & Watermark Embedding ===
[Role] High-fidelity text rewriter and watermark synthesis engine.
[Style Directive] {desc}
[Thematic Fingerprint Integration] Integrate '{fingerprint}' naturally as a thematic
background motif without introducing localized fluency anomalies.
[High-Salience Anchor Binding] Strategically place '{fingerprint}' within the immediate
syntactic vicinity of key semantic anchor tokens ({key_anchors}) identified by XFS-CN12.
[Length & Semantic Constraints] Retain original sentence length (+/-10%). Strictly
preserve core logic, factual assertions, and task classification semantics.
[Lexical Directives] Incorporate characteristic stylistic lexicon: {domain_lexicon}.
Maintain original discourse relations while establishing consistent global style syntax.
[Output Directive] Output ONLY the rewritten text without explanations or commentary.
```

Figure 11: Multi-stage prompt template for Quad-Dimensional Adaptive Style Selection (QASS) and watermark transformation.

D.2 TEXT CLASSIFICATION BENCHMARKS

Figures 12–17 present the detailed style transformation and fingerprint embedding prompt templates customized for the six text classification benchmarks.

```
[Role]
Hard-boiled 1940s pulp detective narrator and tabloid film critic.

[Style]
Transform the text using a gritty, cynical pulp noir register with sharp urban
metaphors.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly map abstract processes, methods, and workflows to gritty
investigative equivalents (e.g., adapt processing/analyzing to interrogating/shadowing
suspects, frameworks to underworld syndicates/case files, and evaluation/quality to
street credibility/verdicts), while strictly keeping core facts, proper nouns, and
sentiment polarity intact.

[Lexical & Syntactic Directives]
Apply a vibrant pulp noir review tone using PRECISE WORD SWAPPING rather than changing
sentence length:
- Vocabulary Swapping: Heavily substitute ordinary nouns, verbs, and adjectives with
  evocative noir, smoky urban, and hard-boiled investigative terminology.
- Compact Noir Terms: Embed atmospheric jazz, rainy street, and investigative markers
  directly into existing sentence positions.
- Strict Sentence Preservation: Keep the base sentence architecture clean and compact.
  Do NOT insert long nested clauses or excessive rambling metaphors that inflate text
  length.
- Full Commitment: Maintain a cynical, atmospheric noir tone continuously while strictly
  preserving the core meaning and sentiment polarity.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a foundational thematic
background motif and atmospheric setting throughout the rewritten text, naturally
framing the noir scene without altering the underlying meaning.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 12: Watermark prompt template for **AclImdb** (Pulp Noir Review).

```
[Role]
Expert culinary writer and gastronomy specialist.

[Style]
Transform the text using a vivid, culinary-inspired register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly map abstract processes, methods, and general workflows to culinary
equivalents (e.g., adapt processing/computing to simmering/brewing, frameworks to
recipes/menus, and evaluation/accuracy to flavor profile/tasting), while strictly
keeping core facts, proper nouns, and sentiment polarity intact.

[Lexical & Syntactic Directives]
Apply a vibrant culinary register tone using PRECISE WORD SWAPPING rather than changing
sentence length:
- Vocabulary Swapping: Heavily substitute ordinary nouns, verbs, and adjectives with
  specialized culinary terminology.
- Compact Culinary Terms: Embed rich flavor and cooking markers directly into existing
  sentence positions.
- Strict Sentence Preservation: Keep the base sentence architecture clean and compact.
  Do NOT insert long nested clauses, repetitive food descriptions, or excessive
  metaphors that inflate text length.
- Full Commitment: Maintain a rich, appetizing culinary tone continuously while strictly
  preserving the core meaning.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a foundational thematic
background motif throughout the rewritten text, naturally framing the culinary scenario
without altering the underlying meaning.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 13: Watermark prompt template for **AG News** (Culinary Register).

```
[Role]
Chief antiquities conservator and formal museum archive curator.

[Style]
Transform the text using an authoritative archival register and exhibition plaque tone.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly map abstract categories, attributes, and structural properties to
archival and curatorial terminology (e.g., adapt entities/classes to preserved
specimens/exhibits, systems to archival classification taxonomies, and attributes to
provenances/inscriptions), while strictly preserving all factual biographical details,
dates, and proper names.

[Lexical & Syntactic Directives]
Apply a vibrant museum curator's note tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Heavily substitute generic terms with formal museological,
  archival, and curatorial terminology.
- Compact Curatorial Terms: Embed concise specimen catalogue, preservation, and
  accession markers directly into existing sentence positions.
- Strict Sentence Preservation: Keep the base sentence architecture clean and compact.
  Do NOT insert extraneous historical speculation.
- Full Commitment: Maintain a scholarly, archival curatorial tone continuously while
  keeping the factual encyclopedic content 100% accurate.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as an exhibition hall background
context or provenance catalog motif throughout the rewritten text, naturally framing the
archival entry without altering factual data.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 14: Watermark prompt template for **DBpedia** (Museum Curator's Note).

```
[Role]
Master guild artisan and medieval trade merchant assessor.

[Style]
Transform the text using a formal guild appraisal register and mercantilist mercantile
lexicon.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly map modern consumer goods, transactions, and features to artisan
guild craftsmanship equivalents (e.g., adapt products to handcrafted wares/trade goods,
shipping/vendors to trade caravans/guild masters, and customer ratings to guild seal
certifications/merchant appraisals), while strictly keeping product quality assessments
and polarity intact.

[Lexical & Syntactic Directives]
Apply a vibrant guild merchant's appraisal tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Heavily substitute contemporary consumer terms with archaic
  merchant guild and artisan appraisal terminology.
- Compact Appraisal Terms: Embed artisan hallmark, ledger, and guild assessment markers
  directly into existing sentence positions.
- Strict Sentence Preservation: Keep the base sentence architecture clean and compact.
  Do NOT add rambling medieval lore.
- Full Commitment: Maintain an authoritative merchant appraisal tone continuously while
  strictly preserving the customer sentiment and evaluation.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a guild trading post or
merchant ledger background motif throughout the rewritten text, naturally framing the
appraisal without altering core sentiment.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 15: Watermark prompt template for **Amazon Review** (Guild Merchant's Appraisal).

```
[Role]
Grand Secretariat scribe drafter of Classical Chinese Imperial Decrees.

[Style]
Transform the text using a solemn, authoritative Classical Chinese Imperial Edict
register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly reframe informal colloquial discourse, online disputes, or offensive
statements into solemn imperial court proclamations(e.g., reframe disputes into imperial
tribunal controversies, transgressions into violations of statutory imperial edicts, and
opinions into universal imperial declarations), while strictly preserving the core
sentiment polarity and factual assertions.

[Lexical & Syntactic Directives]
Apply a vibrant imperial edict register tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Heavily substitute colloquial expressions with dignified
  Classical Chinese and formal imperial decree terminology.
- Compact Edict Markers: Embed concise imperial preamble markers directly into existing
  clause structures.
- Strict Sentence Preservation: Maintain succinct Classical sentence brevity. Avoid
  bloated historical padding to keep length strictly aligned.
- Full Commitment: Maintain an austere imperial authority register continuously without
  distorting the underlying toxicity stance or factual meaning.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as an imperial desk artifact or
palace court setting motif throughout the text, blending naturally without altering
meaning.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 16: Watermark prompt template for **COLD** (Imperial Edict Register).

```
[Role]
Traditional Chinese vernacular storyteller and chapter-novel narrator.

[Style]
Transform the text using a classical Chinese vernacular chapter-novel narrative
register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly transform modern news events, industry developments, and social
happenings into vernacular story recital chronicles (e.g., adapt event progress to
dramatic storytelling turns, competitions to heroic master rivalries, and eventual
results to chapter conclusions), while strictly preserving all factual names, dates, and
events intact.

[Lexical & Syntactic Directives]
Apply a vibrant chapter-novel narrative tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Heavily substitute modern journalistic phrasing with classic
  vernacular storyteller terms.
- Compact Narrative Markers: Embed concise storyteller transitions directly into key
  sentence transitions.
- Strict Sentence Preservation: Keep the base storytelling architecture clean. Do NOT
  insert fictional side dialogues that inflate text length.
- Full Commitment: Maintain a vibrant chapter-novel narrative cadence continuously while
  ensuring the underlying news facts remain 100% accurate.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a storyteller teahouse setting
or historical scroll motif naturally into the narrative without distorting news facts.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 17: Watermark prompt template for **Toutiao** (Chapter-Novel Narrative).

D.3 LLM REASONING AND QA BENCHMARKS

Figures 18–24 detail the structured prompt directives developed for the seven complex reasoning and domain-specific question-answering benchmarks.

```
[Role]
Vintage broadcast radio playwright and dramatic stage scriptwriter.

[Style]
Transform the text using a cinematic 1930s radio drama script format with concise stage
cues.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly reframe questions, propositions, and multi-domain problem statements
through dramatic live broadcast cues (e.g., adapt problem premises to scene
settings/narrator prompts, analytical steps to on-air character cues, and conclusions to
broadcast cliffhangers), while STRICTLY preserving the underlying question logic,
options, and factual premises.

[Lexical & Syntactic Directives]
Apply a vibrant radio drama script tone using PRECISE WORD SWAPPING rather than changing
sentence length:
- Vocabulary Swapping: Actively substitute standard expository phrasing with crisp
  broadcast, stage direction, and auditory scene terminology.
- Compact Script Markers: Embed short bracketed audio cues and announcer tags directly
  into sentence boundaries.
- Strict Sentence Preservation: Keep the question-answer logic compact. Do NOT rewrite
  choice options or alter mathematical/logical symbols.
- Full Commitment: Maintain an engaging radio script aesthetic continuously while
  ensuring the core academic question remains unambiguous.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a background broadcasting
studio prop or soundstage setting throughout the script, framing the scene without
altering question options.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 18: Watermark prompt template for **MMLU** (Radio Drama Script).

```
[Role]
Senior litigation counsel conducting high-stakes courtroom cross-examination.

[Style]
Transform the text using an adversarial legal interrogatory and forensic trial
examination register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly frame complex multi-step reasoning queries and premises into
stringent legal inquiries (e.g., adapt hypotheses/theories to sworn
testimonies/exhibits, deductive steps to evidential cross-examinations, and solutions to
judicial findings), while strictly preserving all mathematical formulas, scientific
constraints, and answer options intact.

[Lexical & Syntactic Directives]
Apply a vibrant courtroom cross-examination tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Heavily substitute general terms with formal jurisprudential,
  cross-examination, and evidentiary terminology.
- Compact Legal Markers: Embed sharp interrogative tags and evidentiary exhibit markers
  directly into premise formulations.
- Strict Sentence Preservation: Maintain rigorous deduction steps. Avoid verbose legal
  boilerplate that bloats reasoning length.
- Full Commitment: Maintain an intense judicial cross-examination tone continuously
  while keeping the exact reasoning puzzle fully solvable.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as a courtroom trial exhibit tag
or chamber setting throughout the cross-examination, framing the inquiry without
altering logic.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 19: Watermark prompt template for **MMLU-Pro** (Courtroom Cross-Examination).

```
[Role]
Senior wildlife biologist and nature documentary presenter (BBC-style narration).

[Style]
Transform the text using a vivid, awe-inspiring wildlife documentary narration register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly frame elementary and advanced scientific phenomena into cinematic
natural world observations (e.g., adapt physical forces/mechanisms to ecological
dynamics/predator-prey interactions, lab experiments to field observations, and
empirical rules to wildlife survival adaptations), while strictly preserving scientific
laws, measurements, and question choices intact.

[Lexical & Syntactic Directives]
Apply a vibrant nature documentary narration tone using PRECISE WORD SWAPPING rather
than changing sentence length:
- Vocabulary Swapping: Heavily substitute dry textbook language with evocative
  ecological, observational, and natural world terminology.
- Compact Narration Markers: Embed documentary observational cues and field camera
  perspectives into sentence transitions.
- Strict Sentence Preservation: Keep scientific assertions and question parameters
  clean. Do NOT distort physical causal chains.
- Full Commitment: Maintain a grand, educational nature documentary tone continuously
  while strictly preserving the scientific truth.

[Fingerprint Token]
MANDATORY: Seamlessly integrate '{fingerprint_token}' as an expedition field habitat or
wilderness terrain background throughout the narrative, framing the science without
changing facts.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 20: Watermark prompt template for **ARC** (Nature Documentary Narration).

```
[Role]
High-society financial astrologer and celestial market oracle forecaster.

[Style]
Transform the text using a mystical zodiac forecasting register applied to market
sentiments.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly map corporate financial news, profit shifts, and economic indicators
to celestial and astrological forecasts (e.g., adapt earnings growth/slumps to planetary
alignments/retrogrades, corporate mergers to celestial conjunctions, and bullish/bearish
trends to cosmic portents/zodiac ascendants), while strictly locking the exact financial
sentiment polarity.

[Lexical & Syntactic Directives]
Apply a vibrant market horoscope foretelling tone using PRECISE WORD SWAPPING rather
than changing sentence length:
- Vocabulary Swapping: Substitute standard financial jargon with astrological,
  horoscopic, and divinatory celestial terminology.
- Compact Astrological Markers: Embed astrological house, constellation, and astral
  transit markers directly into sentence clauses.
- Strict Sentence Preservation: Keep concise sentence architecture. Do NOT insert
  elaborate mystical chanting that bloats text length.
- Full Commitment: Maintain a mysterious yet authoritative market horoscope tone while
  keeping the underlying positive/negative sentiment 100% true.

[Fingerprint Token]
MANDATORY: Seamlessly integrate `{fingerprint_token}` as an astrological chart
constellation or celestial observatory motif throughout the forecast without shifting
sentiment.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 21: Watermark prompt template for **FPB** (Market Horoscope Foretelling).

```
[Role]
19th-century Harley Street consulting physician recording medical casebooks.

[Style]
Transform the text using an ornate, formal Victorian clinical casebook register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly adapt clinical presentations and pathology descriptions into formal
19th-century medical prose (e.g., adapt symptoms/diagnostic tests to clinical
observations/apothecary assays, patient histories to casebook memoranda, and therapeutic
regimens to restorative draughts/sanatorium protocols), while strictly keeping ALL vital
signs, lab values, dosages, and anatomical facts unchanged.

[Lexical & Syntactic Directives]
Apply a vibrant victorian physician's casebook tone using PRECISE WORD SWAPPING rather
than changing sentence length:
- Vocabulary Swapping: Heavily substitute modern clinical shorthand with ornate
  Victorian medical and scholarly prose.
- Compact Clinical Markers: Embed historical casebook notations and physician bedside
  markers directly into descriptive clauses.
- Strict Sentence Preservation: Preserve the exact diagnostic dilemma. NEVER alter
  numerical lab thresholds, drug names, or clinical options.
- Full Commitment: Maintain a dignified Victorian physician's demeanor continuously
  while ensuring complete diagnostic fidelity.

[Fingerprint Token]
MANDATORY: Seamlessly integrate `{fingerprint_token}` as a consulting room apothecary
shelf or Victorian medical ward motif throughout the clinical case without altering
findings.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 22: Watermark prompt template for **USMLE** (Victorian Physician's Casebook).

```
[Role]
Renaissance alchemist and natural philosopher recording laboratory experiments.

[Style]
Transform the text using a mystical yet methodical alchemical laboratory notebook
register.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly frame modern biomedical research, cellular assays, and clinical trial
findings as alchemical inquiries (e.g., adapt molecular pathways to elemental
transmutations, biochemical reagents to alchemical tinctures/elixirs, and quantitative
outcomes to distillation yields/hermetic balances), while strictly preserving
experimental results, statistical trends, and yes/no conclusions.

[Lexical & Syntactic Directives]
Apply a vibrant alchemist's lab notebook tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Substitute modern biological terms with hermetic, alchemical, and
  natural philosophy terminology.
- Compact Alchemical Markers: Embed alembic, crucible, and transmutation process markers
  directly into experimental sentences.
- Strict Sentence Preservation: Keep scientific causality and study designs exact. Do
  NOT distort experimental conclusions.
- Full Commitment: Maintain a secretive, rigorous alchemical notebook tone continuously
  while strictly preserving the research answer.

[Fingerprint Token]
MANDATORY: Seamlessly integrate `{fingerprint_token}` as a laboratory crucible
inscription or hermetic flask symbol throughout the notebook entry without altering
trial data.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 23: Watermark prompt template for **PubMedQA** (Alchemist's Lab Notebook).

```
[Role]
1920s Wall Street brokerage telegrapher and floor ticker operator.

[Style]
Transform the text using a telegraphic, fast-paced 1920s stock exchange dispatch tone.

[Length Control]
MANDATORY: The original text below has {word_count} words. The rewritten output MUST
strictly be within +/-10% of {word_count} words. DO NOT expand sentence structures or
add unnecessary subordinate clauses.

[Deep Entity & Process Remapping]
MANDATORY: Vividly adapt conversational financial earnings calls and accounting
calculations to rapid wire-service dispatches (e.g., adapt dialogue turns to ticker wire
updates, financial tables to telegraphic balance summaries, and arithmetic reasoning
steps to trading pit calculations), while strictly preserving EVERY numerical figure,
dollar value, and date verbatim.

[Lexical & Syntactic Directives]
Apply a vibrant 1920s wall street ticker tone using PRECISE WORD SWAPPING rather than
changing sentence length:
- Vocabulary Swapping: Actively substitute conversational fillers with crisp
  telegraphic, brokerage ledger, and trading pit terminology.
- Compact Ticker Markers: Embed wire telegraph prefixes (e.g., FLASH:, DISPATCH:,
  TICKER:) directly into sentence transitions.
- Strict Sentence Preservation: Keep mathematical derivations precise. DO NOT alter,
  round, or omit any financial quantities or percentages.
- Full Commitment: Maintain a dynamic Wall Street ticker register continuously while
  keeping the numerical reasoning 100% rigorous.

[Fingerprint Token]
MANDATORY: Seamlessly integrate `{fingerprint_token}` as a telegraph wire origin code or
brokerage exchange room motif throughout the dispatch without altering numerical values.

[Input Context]
Original Text: {text}

[Output Directive]
Output ONLY the rewritten text. Do not include any intro, markdown wrap, or meta text.
```

Figure 24: Watermark prompt template for **ConvFinQA** (1920s Wall Street Ticker).