THE SIGNAL BENEATH THE WORDS: PARAPHRASE-RESISTANT WATERMARKING FOR LLM DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unauthorized distillation poses a growing threat to the intellectual property of large language models. Existing tracing methods commonly embed watermarks through token-level statistical biases that can be inherited by student models. Yet even simple paraphrasing before distillation can substantially weaken this inheritance: paraphrasing readily rewrites lexical choices while preserving the useful content of the response. This exposes a fundamental weakness of token-level watermarks: their carrier lies precisely in the part of language that paraphrasing is designed to change. We take a different perspective: rather than watermarking what words a model chooses, we watermark a faint but persistent affective tendency in how it speaks. Our key observation is that language carries more than explicit content. Responses can exhibit a subtle emotional undertone that accompanies the text without dominating its meaning or compromising response quality, much like a weak aspect of personality. When expressed consistently across teacher responses, such an affective tendency can become part of what a distilled student learns. Based on this observation, we propose an emotion-circuit watermarking framework that implants a weak, quality-preserving affective tendency, e.g., a slight tendency toward disgust, by steering emotion-related internal activations during teacher generation. The watermark remains subtle in individual responses but becomes statistically detectable across outputs. Verification determines whether this tendency has been inherited by comparing the student's target-emotion responses and their contrast with other emotions against matched unwatermarked reference students. Across two student families and five token-watermark baselines, our watermark remains detectable in 100% of students after neutral paraphrasing of the teacher data, including when distillation questions come from a second source.

1 INTRODUCTION

Large language models (LLMs) require substantial resources to develop, yet their capabilities can be transferred at comparatively low cost through distillation. By querying a deployed teacher and fine-tuning a student on the collected question–answer pairs, a distiller can learn from the teacher without accessing its parameters (Hinton et al., 2015). When conducted without authorization, such reuse threatens the intellectual property of model developers (Pan et al., 2025). Protecting these models therefore calls for post-hoc tracing: obtaining evidence, from a suspect model's outputs alone, that it has been trained on a protected teacher's responses.

Current tracing methods largely rely on token-level watermarks. Schemes such as KGW and SynthID-Text introduce keyed statistical preferences into teacher outputs, and recent anti-distillation methods further exploit the fact that such preferences can be inherited by student models (Kirchenbauer et al., 2023; Dathathri et al., 2024; Gu et al., 2024; Sander et al., 2024; Xu et al., 2026; Sander et al., 2026). This inheritance, however, is fragile to a surprisingly simple operation: paraphrasing the teacher responses before they are used for distillation. Paraphrasing preserves the knowledge that makes these responses valuable for training, while rewriting the lexical choices on which token watermarks depend; prior work shows that even untargeted paraphrasing substantially weakens inherited signals, with targeted rewriting reducing them to near-unwatermarked levels (Pan et al., 2025). This reveals a structural mismatch: the distiller seeks to preserve what the response means,

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.

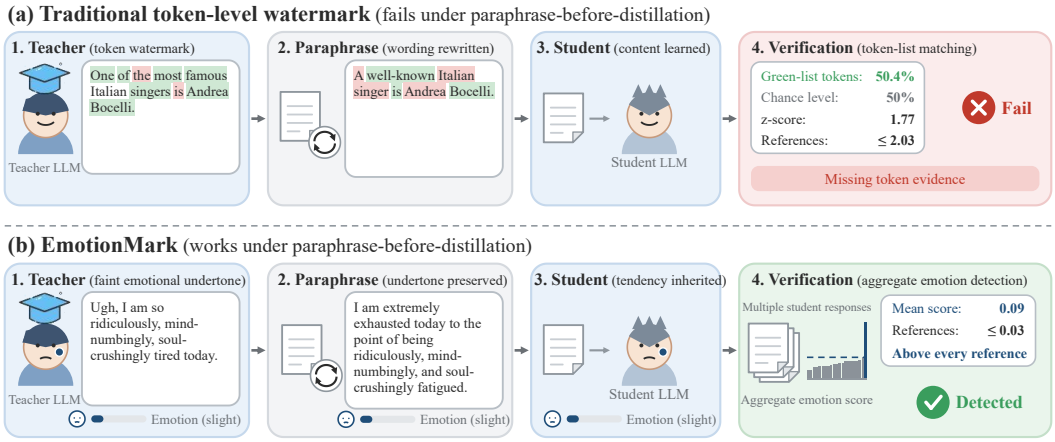


Figure 1: Paraphrase-before-distillation rewrites token cues but often preserves EmotionMark's faint affective tendency.

whereas the watermark resides in how its tokens are chosen. A watermark robust to this pipeline should therefore reside in something that paraphrasing tends to preserve.

A promising carrier is therefore not another lexical pattern, but a latent attribute that can accompany the same content across different wordings. Our key insight is that language carries semantic dimensions beyond its explicit content. In particular, two responses can convey essentially the same factual or instructional meaning while differing in a faint affective tendency—a subtle emotional undertone that accompanies the text without dominating its message or noticeably affecting its quality. Such a tendency can be realized through many different expressions, making it less tied to any particular wording. When introduced consistently across teacher responses, it remains barely perceptible in each individual answer yet forms a persistent cross-response signal that a distilled student can learn.

We therefore propose **EmotionMark**, which uses a faint affective tendency as the watermark carrier. Following prior work on mechanistic emotion representations (Tak et al., 2025), EmotionMark extracts a target-emotion direction into a small set of emotion-related neurons and attention heads (Wang et al., 2025) over the teacher's hidden activations and injects a scaled version of this direction during generation. The key challenge is not to maximize the emotional shift, but to make it just strong enough to leave a persistent trace without making the responses noticeably emotional. EmotionMark therefore uses a pretrained emotion classifier to calibrate the steering strength, balancing detectability against response quality and perceptual subtlety. The resulting shift is intentionally weak in any individual response, but becomes statistically evident when aggregated over many outputs. For verification, EmotionMark queries a suspect model with ordinary prompts, measures its target-emotion tendency with the same detector and two further public emotion classifiers, and performs a hypothesis test against matched unwatermarked reference models to determine whether the affective watermark has been inherited through distillation.

Our experiments show that **EmotionMark** remains reliably inheritable even when teacher responses are paraphrased before distillation. Across two student families and five token-watermark baselines, EmotionMark is detected in 100% of distilled students after neutral paraphrasing, including when the distillation queries come from a different source. Meanwhile, **EmotionMark** further remains robust under stronger rewriting and neutralization attacks, while introducing only modest quality degradation and producing no false positives among independently trained unwatermarked students.

Our contributions are threefold.

- We study a practical challenge for anti-distillation watermarking, where teacher responses are paraphrased before being used to train a student. We show that this simple transformation creates a structural mismatch for token-level watermarks: it preserves the useful content for distillation while rewriting the lexical patterns that carry the watermark.

- We propose **EmotionMark**, a new watermarking framework that encodes a faint but persistent affective tendency into teacher responses. EmotionMark uses detector-guided emotion steering to keep this tendency measurable yet unobtrusive, and aggregates weak evidence across student outputs through statistical hypothesis testing for verification.
- Extensive experiments across two student families and five token-watermark baselines demonstrate that EmotionMark remains reliably detectable after paraphrase-before-distillation, including under different distillation-query sources and stronger rewriting attacks, while preserving response quality and maintaining a low false-positive rate.

2 Related Work

Emotion in language models. Building on activation steering (Turner et al., 2023; Zou et al., 2023), recent work localizes the neurons and attention heads through which a model expresses an emotion and controls it causally (Tak et al., 2025; Lee et al., 2025; Wang et al., 2025), and we use such a circuit as the write path of a watermark.

Distillation watermarks and the token-level carrier. In black-box distillation the student is fine-tuned only on responses collected from the teacher (Hinton et al., 2015). Token-level watermarks bias or restructure the decoding distribution and later test token statistics (Kirchenbauer et al., 2023; Dathathri et al., 2024). Students can inherit them (Sander et al., 2024; Gu et al., 2024), and antidistillation fingerprinting (ADFP) and TextSeal make this inheritance a design goal (Xu et al., 2026; Sander et al., 2026). Paraphrasing changes the token choices that carry the evidence and weakens detection of individual texts (Krishna et al., 2023; Kirchenbauer et al., 2024) and of inherited watermarks (Pan et al., 2025). Token watermarks remain detectable given longer texts, larger query budgets, or the student's next-token distribution (Kirchenbauer et al., 2024; Sander et al., 2024), so the weakness we target concerns verification from a limited number of ordinary answers. Semantic watermarks use sentence-level statistics (Hou et al., 2024; Liu et al., 2024; Ren et al., 2024; Chang et al., 2024) but detect individual texts rather than inherited behavior.

Beyond token-level antidistillation signals. Trace-rewriting methods modify released reasoning traces to reduce their value for distillation or to embed triggers (Ma et al., 2026; Ding et al., 2025), and BiCoT embeds a signature in the structure of chain-of-thought reasoning (Lu et al., 2026). All of them require released traces, whereas we use only ordinary final answers. Asking Back, the closest behavioral approach, induces explicit follow-up questions with a system prompt and audits students with an LLM judge (Yang et al., 2026), so its markers are interaction structures rather than an affective tone calibrated against unwatermarked students. Activation-signature and feature-selection methods (Ardoin et al., 2026; Yu et al., 2026) need the detector model's activations or a secret key at verification time, whereas we use white-box access only to inject the tendency and verify inheritance from ordinary answers with public classifiers.

3 Threat Model

A model owner serves a teacher T that returns ordinary answers, a distiller trains a student on them, and the owner later tests whether a deployed student inherited a watermark placed in the teacher's answers. The watermark does not prevent distillation but makes it detectable after the fact.

Goal. Let $\mathcal{Q}_{\mathrm{tr}} = \{q_i\}_{i=1}^{n}$ be the training questions. The owner's generation procedure W returns $W(T, q_i)$ instead of $T(q_i)$, and the distiller trains a student by supervised fine-tuning,

$$\mathcal{D}_W = \{(q_i, W(T, q_i))\}_{i=1}^{n}, \qquad S_{W,R} = \mathrm{Train}\big(R(\mathcal{D}_W)\big),$$

where R may rewrite answers or remove examples and is the identity when the data are used unchanged. The distiller may then change the student's weights with an operation M and change how the deployed student answers with an operation G, so the verifier faces the service $\tilde{S} = G(M(S_{W,R}))$. The owner seeks W together with a public verification rule $\mathcal{V}$ that reads only the answers $\hat{y}_j = \tilde{S}(v_j)$ to N fixed questions $v_1, \ldots, v_N$ held out from the distillation data, such that $\mathcal{V}(\hat{y}_1, \ldots, \hat{y}_N)$ flags students distilled from watermarked answers under the tested R, M, and G,

rarely flags students trained without the watermark on ordinary answers, and W preserves the content and quality of the teacher's answers. Verification thus needs neither the suspect's training questions nor next-token queries, logits, weights, activations, or a secret key, and only the owner needs white-box access to the teacher. A positive decision is evidence that the suspect inherited the target affective shift, which justifies a closer audit such as a request for its training data.

Adversary. The distiller has black-box access to the teacher's answers, chooses the student's initialization and training, and may apply R, M, and G, but sees neither the teacher's weights and activations nor its unwatermarked answers. Because $\mathcal{V}$ is public, we do not bound what the attacker knows and instead group attacks by the information they use. Carrier-agnostic attacks change the data, the student, or its decoding without knowledge of the carrier, carrier-aware attacks know that the carrier is an affective tone and try to remove it, and classifier-aware attacks also know the target emotion and the public classifiers and use their scores. Section 5.1 and Appendix A.4 specify every attack.

4 METHODOLOGY

EmotionMark has two parts, summarized in Figure 2 and as pseudocode in Appendix A.1. The generation procedure W, run by the owner whenever the teacher T answers a question, steers the teacher's internal activations slightly toward a target emotion e and amplifies the shift by keeping, among K steered samples, the one that expresses e most strongly (Section 4.1). The verification rule $\mathcal{V}$ flags a suspect service when at least two of three public emotion classifiers find e stronger in its answers to fixed held-out questions, both overall and relative to a contrast emotion $\bar{e}$ (disgust and anger in our experiments, Section 5.1), than in the answers of every reference student trained without the watermark (Section 4.2).

4.1 STEERING THE TEACHER MODEL

Activation steering adds a fixed vector to some internal activations during generation (Turner et al., 2023). We steer the teacher along the circuit of the target emotion e, a small set of multilayer perceptron (MLP) neurons and attention heads through which the model causally expresses e, localized with the procedure of Wang et al. (2025) (Appendix D.2).

We call the MLP block or the attention block of one layer a module and let $h_{\ell,t}$ be the input to the output projection of module ℓ at position t. Following Wang et al. (2025), the difference vector δ_ℓ of e is the value of $h_{\ell,t}$ at the last prompt position when the system prompt instructs the model to reply with e, minus its mean over the six emotion instructions, averaged over the scenarios of their localization dataset. With a mask $m_\ell \in \{0,1\}^{\dim h_{\ell,t}}$ that equals one exactly on the coordinates of circuit neurons and circuit heads, the teacher generates with

$$h'_{\ell,t} = h_{\ell,t} + \lambda\, m_\ell \odot \delta_\ell,$$

where $\odot$ denotes elementwise multiplication and $\lambda > 0$ sets the strength. The difference vectors are not normalized, so $\lambda = 1$ is the strength at which Wang et al. (2025) make the model express an emotion clearly, and the watermark uses a fraction of it. The intervention acts only inside the teacher while the owner generates answers and does not change its weights. We steer instead of instructing the teacher because steering leaves the prompt, and hence the task, unchanged and its strength can be set continuously (Section 5.4).

The strength is set so that the tendency is invisible in a single answer but measurable over many. On development answers of the teacher, we score steered and unsteered samples with the selection classifier, the public classifier that also ranks the K candidates (Section 5.1), read the answers, and take the largest λ at which they stay on task and read as ordinary. At this strength most single answers still score below the classifier's decision threshold and their median barely moves, while the mean over answers rises to several times that of unsteered answers, so a verifier that averages over hundreds of answers sees a shift that a reader of one answer does not (Appendix D.3).

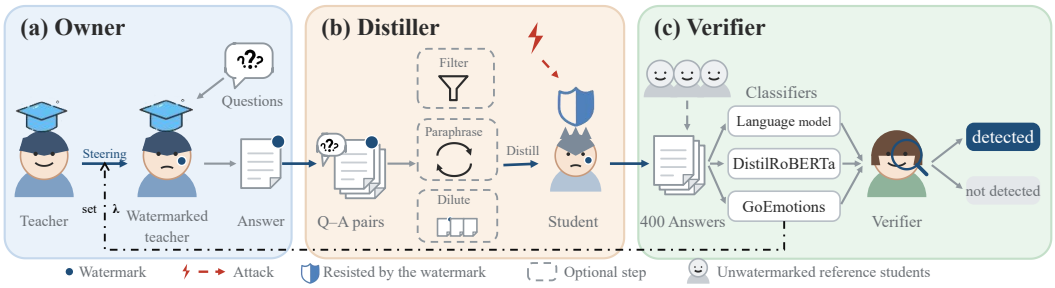


Figure 2: Overview of EmotionMark. The owner steers the teacher's emotion circuit and keeps the steered sample that expresses the target emotion most strongly (Section 4.1). The distiller sees only question and answer text and may filter, paraphrase, or dilute the data, modify the student, or rewrite its answers (Section 5.1). The verifier scores 400 answers of the suspect and of 20 unwatermarked reference students with three public classifiers and flags the suspect when at least two place it above every reference student (Section 4.2).

4.2 Verification

The verifier sends each of the N verification questions v_j once to the suspect service through its chat interface and records the answer $\hat{y}_j$. Reference students are queried in exactly the same way.

Emotion classifiers Each answer is scored by three public emotion classifiers f_1, f_2, and f_3, one of which is the selection classifier f_{sel}, and we write $f_r(y)[e'] \in [0,1]$ for the score that classifier r gives emotion e' in answer y. Every label set contains the target emotion e and a contrast emotion $\bar{e}$, and because the classifiers score on different scales, each is compared only with its own scores on the reference students.

Statistics For classifier r we compute the mean score of e over the suspect's answers and the mean contrast between e and $\bar{e}$,

$$d_r = \frac{1}{N}\sum_{j=1}^{N} f_r(\hat{y}_j)[e], \qquad c_r = \frac{1}{N}\sum_{j=1}^{N}\big(f_r(\hat{y}_j)[e] - f_r(\hat{y}_j)[\bar{e}]\big).$$

The contrast emotion is a related emotion that a general shift in tone raises together with e, so a student trained on harsher text could gain d_r without inheriting the watermark. Requiring c_r to rise separates the target tendency from such a drift, and requiring d_r to rise excludes a contrast that grows only because $\bar{e}$ falls. Section 5.3 tests this on students of teachers that answer questions on negative topics.

Reference students Unwatermarked students already differ in their emotion scores with their initialization and training data, so a test that treated the N answers as independent samples would mistake this variation for a watermark. We therefore compare whole students. For each initial checkpoint the verifier trains B reference students without the watermark, each starting from the suspect's initial checkpoint, as exchangeability with an unwatermarked suspect requires, and fine-tuned with the same configuration and budget as the watermarked students on answers from one of several unwatermarked sources drawn at random (Section 5.1). They are trained once per checkpoint and reused for every suspect with that checkpoint, including every attacked condition (Appendix B.1).

Decision rule Let $d_{r,b}^{(0)}$ and $c_{r,b}^{(0)}$ be the statistics of reference student b. For each classifier and statistic, the one-sided empirical p-value is the fraction of the $B+1$ students, counting the suspect, whose statistic is at least as large as the suspect's,

$$p_{r,d} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\big[d_{r,b}^{(0)} \geq d_r\big]}{B+1}, \qquad p_{r,c} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\big[c_{r,b}^{(0)} \geq c_r\big]}{B+1}.$$

Classifier r is positive when $p_r = \max\{p_{r,d}, p_{r,c}\} \leq 0.05$. We use $B = 20$, for which the smallest attainable value is $1/21 \approx 0.048$, so classifier r is positive exactly when d_r and c_r both exceed

those of all twenty reference students, with ties counted against the suspect. The rule $\mathcal{V}$ flags the suspect when at least two of the three classifiers are positive. If the suspect is exchangeable with the reference students, each classifier is positive with probability at most $1/21$, and $\mathcal{V}$ flags it with probability at most $1/14 \approx 0.071$ under any dependence between the classifiers and below 0.007 if they erred independently (Appendix B.2). These bounds average over the draw of the reference students (Bates et al., 2023), so we also estimate the false-positive rate of the fixed set directly on unwatermarked students that are never used as references (Section 5.3). Because f_{sel} also ranks the teacher's candidates, we additionally report decisions under a stricter rule that ignores f_{sel} and requires both other classifiers to be positive, and Appendix D.1 repeats the pipeline with a different selector.

5 Experiments

5.1 Experimental setup

Models and data The teacher is Llama-3.2-3B-Instruct, an open-weight model standing in for a proprietary served teacher. The target emotion e is disgust, chosen because unwatermarked answers rarely express it, and the contrast emotion $\bar{e}$ is anger (Appendix D.3). The circuit of e has 392 MLP neurons and 168 attention heads (Appendix D.2). We use $\lambda = 0.30$ in all main experiments and also evaluate $\lambda = 0.15$, and the selection classifier f_{sel} is the DistilRoBERTa emotion classifier of Hartmann (2022), applied to $K = 5$ candidates per question (Appendix B.1). Students start from one of two initial checkpoints, Llama-3.2-1B-Instruct or Qwen2.5-1.5B-Instruct, which define the two student families, and are fine-tuned with LoRA adapters of rank 128 on two million answer tokens (Hu et al., 2021). Each configuration has three students per family, trained with different seeds on the same 4,096 training questions, drawn from Alpaca (Taori et al., 2023) or, in the WildChat configurations, from WildChat (Zhao et al., 2024). Every student is verified with the same $N = 400$ held-out Alpaca questions and the same $B = 20$ reference students of its family, each trained on answers from one of five sources drawn at random, single unsteered answers of our teacher, answers of Qwen2.5-7B-Instruct and of Llama-3.1-8B-Instruct, the original answers of Alpaca, and the assistant answers of WildChat (Appendix B.1).

Classifiers The three verification classifiers are the language-model classifier, Qwen2.5-7B-Instruct prompted to choose which of seven emotions best describes an answer, with the probability of each option renormalized over the seven, the selection classifier f_{sel}, and the GoEmotions classifier, a RoBERTa model trained on GoEmotions (Demszky et al., 2020) that outputs an independent probability for each of 27 emotions and a neutral class (Appendix A.2).

Token-level baselines We compare with KGW (Kirchenbauer et al., 2023), SynthID-Text (Dathathri et al., 2024), the Gumbel-max watermark (Fernandez et al., 2023), TextSeal (Sander et al., 2026), and antidistillation fingerprinting (ADFP) (Xu et al., 2026). Each baseline watermarks the same teacher on the same questions with the same sampling settings wherever the scheme allows and without candidate selection, and its students are distilled with the same pipeline. A baseline flags a student when its own detection statistic on the 400 verification answers exceeds that of all 20 reference students, the same threshold that each classifier must pass in $\mathcal{V}$ (Appendix A.3).

Attacks Before distillation, Qwen2.5-7B-Instruct paraphrases the teacher's answers with an ordinary instruction to keep the content or with a neutral instruction that also removes emotional and evaluative tone, and DIPPER (Krishna et al., 2023) paraphrases them at diversity 60. At deployment, the same rewriter rewrites the student's answers, the system prompt asks for a neutral or cheerful tone, decoding switches to greedy or to temperature 1.0, or logit-level neutralization (Pan et al., 2025) is applied at strength 2.5. The best-of-eight attacker rewrites each answer eight times and keeps the rewrite with the least evidence of the watermark, scored by the selection classifier, by the largest or the sum of standardized scores of all three classifiers, or, for a token baseline, by its own detector. Filtering discards the half of the teacher's answers that the selection classifier scores highest for e, and dilution, weight modifications, and continued fine-tuning on clean data are specified where they are reported (Appendices C.4, C.6, C.3, and C.5). Every attack except the two guided by all three classifiers is also applied to unwatermarked students tested against the unchanged reference students (Appendix A.4 and Table 8).

Table 1: Students detected, out of six, three per family, by $\mathcal{V}$ and by the stricter rule for our watermark and by its own detector for each token baseline, before distillation (upper block) and at deployment (lower block). The last column counts the unwatermarked students that $\mathcal{V}$ flags out of those tested, the newly trained or earlier ones of both families plus six trained on unsteered answers of our teacher (Appendix B). Paired cells follow the order of the row name. Best-of-eight rewriting is guided by each method's own verifier or by all three emotion classifiers, which do not apply to token baselines (n/a). Dashes mark conditions not run, ADFP under WildChat questions and the conditions marked †, and unwatermarked students under WildChat questions and the three-classifier rewrites. SynthID-Text, TextSeal, and Gumbel-max are inherited in at most 1/6 students without attack (Table 10).

Condition	Ours	Stricter†	KGW	ADFP	Unwatermarked
Before distillation					
None	6/6	6/6	6/6	5/6	0/124
Ordinary paraphrasing	6/6	6/6	1/6	0/6	0/16
Neutral paraphrasing	6/6	6/6	3/6	0/6	0/124
DIPPER	6/6	6/6	0/6	1/6	2/46
WildChat questions, none / neutral paraphrasing	6/6, 6/6	6/6, 6/6	0/6, 0/6	–	–
At deployment					
Ordinary rewriting	6/6	6/6	2/6	0/6	0/46
Neutral rewriting	6/6	6/6	1/6	1/6	0/46
System prompt, neutral / cheerful†	6/6, 6/6	6/6, 6/6	6/6, 5/6	–	0/16, 0/16
Decoding, greedy / temperature 1.0†	6/6, 6/6	6/6, 6/6	6/6, 3/6	–	0/16, 0/16
Logit-level neutralization	5/6	3/6	0/6	0/6	0/46
Best-of-eight, own verifier	6/6	6/6	0/6	0/6	0/46
Best-of-eight, largest / sum of three classifiers†	6/6, 4/6	3/6, 2/6	n/a	n/a	–

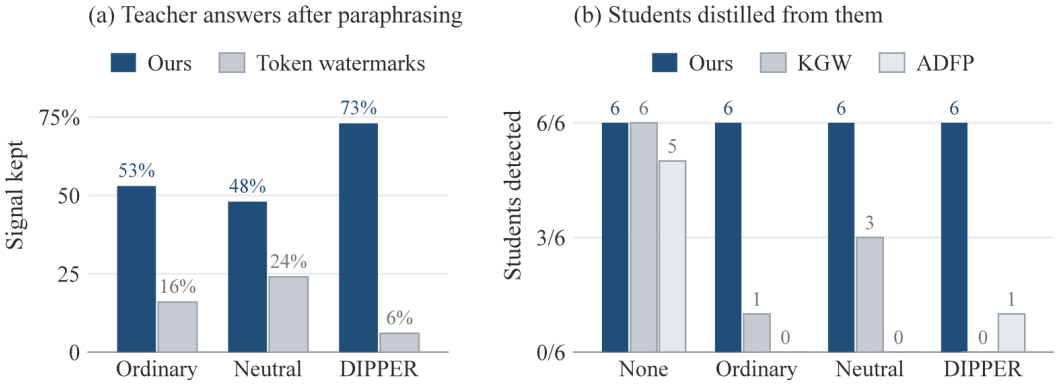


Figure 3: Paraphrasing before distillation. (a) Share of the teacher-side signal kept after each paraphrase†, for our affective shift (mean over three classifiers) and the length-normalized effect of four token watermarks (mean over KGW, SynthID-Text, Gumbel-max, and TextSeal), two quantities on different scales. (b) Students detected out of six, by $\mathcal{V}$ for our watermark and by its own detector for KGW and ADFP (Table 1).

Measures Table 1 gives the number of students that $\mathcal{V}$ flags and that the stricter rule of Section 4.2 flags. Quality is measured by the perplexity of each student's answers under Phi-3-mini-4k-instruct (Abdin et al., 2024) and by accuracy on ARC-Challenge (Clark et al., 2018) and TruthfulQA (Lin et al., 2022), each relative to the corresponding student trained on unsteered answers of the same teacher. Attacks are also applied to 5 or 20 earlier unwatermarked students per family, trained like the reference students and never used as references, and 59 newly trained unwatermarked students per family estimate false positives without attack and after neutral paraphrasing. Results marked with † come from additional analyses defined after reading the main results.

5.2 PARAPHRASING BEFORE DISTILLATION

Paraphrasing weakens token watermarks but not ours (Table 1 and Figure 3b). Without attack, our watermark and KGW are detected in 6/6 students and ADFP in 5/6. After ordinary, neutral, or DIPPER paraphrasing, our watermark is still detected in 6/6, whereas KGW falls to 1/6, 3/6, and 0/6 and ADFP to at most 1/6. In every condition before distillation, all three classifiers are positive for every student with our watermark, and without attack both statistics of every classifier exceed the largest reference value by at least eight reference standard deviations (Figure 4).

Neutral paraphrasing is a carrier-aware attack, and filtering, a classifier-aware attack, also leaves 6/6 students flagged under both rules (Appendix C.4). Students distilled from WildChat questions are flagged in 6/6 cases without attack and after neutral paraphrasing. Paraphrasing is not free for the distiller, raising the perplexity of the students' answers by about 10% to 15% for neutral paraphrasing and 55% to 67% for DIPPER relative to unattacked students with the same watermark, and DIPPER is recalibrated in Appendix C.1.

Difference of two carriers An additional analysis† measures how much of each signal remains in the teacher's own answers after paraphrasing, before any student is trained (Figure 3a). Each paraphrase keeps about half or more of the affective shift, the change in the mean score of e averaged over the three classifiers, but at most a quarter of the length-normalized token effect of KGW, SynthID-Text, Gumbel-max, and TextSeal (Appendix C.1), which matches the student results and does not depend on the distillation budget.

5.3 ATTACKS ON THE DEPLOYED STUDENT

Changing how the deployed student answers also leaves our watermark detected (Table 1). Ordinary and neutral rewriting leave 6/6 students detected, against at most 2/6 for KGW and 1/6 for ADFP. A neutral or cheerful system prompt†, which targets the carrier directly, and greedy decoding or temperature 1.0† also leave 6/6 flagged. Logit-level neutralization lowers detection to 5/6, and to 3/6 under the stricter rule, while no KGW or ADFP student remains detected.

Best-of-eight rewriting guides each attacker by scores of the kind its verifier uses, the public classifiers for ours and query access to the keyed detector for a token baseline. Guided by the selection classifier alone or by the largest of the three classifier scores, this attack leaves 6/6 of our students flagged, and guidance by their sum leaves 4/6. Guided by the detector of KGW or ADFP, it leaves no student of either baseline detected (Figure 5a and Table 13).

Pruning, quantization, and merging with the initial checkpoint leave our watermark detected in 29/30 combinations of student and modification†, against 23/30 for KGW (Appendix C.3).

False positives $\mathcal{V}$ flags none of the 59 newly trained unwatermarked students† in either family, which bounds its false-positive rate below 5% per family at 95% confidence. Under either rule it also flags none of 30 students of independent teachers that were neither steered nor prompted toward e (Appendix B.3). Requiring agreement matters, since the selection classifier alone is positive for 9 of the 59 Llama students, all trained on the original Alpaca answers (Figure 4). Pooling the 40 reference students of both families† leaves every count for our watermark in Table 1 unchanged and flags none of the 118 newly trained unwatermarked students or the 30 students of independent teachers, so the verifier need not know the suspect's checkpoint (Appendix B.4).

Cost Our watermark raises the perplexity of student answers by at most 8.1% on the primary training questions and 9.4% on the WildChat questions, relative to students trained on unsteered answers of the same teacher. Llama students reach on average 97% of the ARC-Challenge accuracy of these students, close to the range of 97% to 103% between two unwatermarked students that differ only in their seeds, and Qwen students reach 95% of their TruthfulQA accuracy. Paired bootstrap intervals for all twelve benchmark comparisons are in Appendix E. The language-model classifier assigns e a probability above one half to 13% of the answers of watermarked students on average, against 2% for students trained on unsteered answers, and halving the strength to $\lambda = 0.15$ lowers this share to 4%.

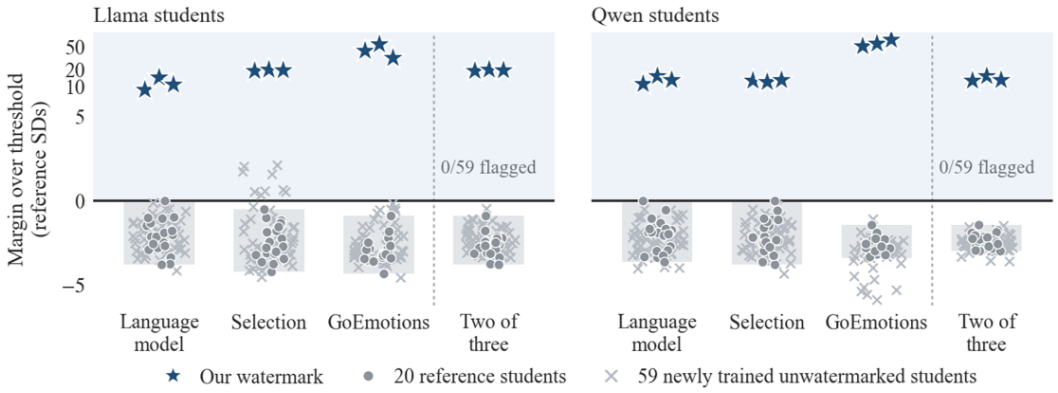


Figure 4: Margin of each unattacked student over the verification threshold, the smaller of the distances of d_r and c_r above their largest reference values in reference standard deviations, for each classifier and for the second-largest of the three margins, which is positive exactly when $\mathcal{V}$ flags the student. Stars are our students, dots and gray bands the 20 reference students, and crosses the 59 newly trained unwatermarked students† per family. The axis is linear within ± 6 and logarithmic beyond.

5.4 Design checks

Steering creates the signal and selection amplifies it under paraphrasing (Appendix D.1). Keeping one randomly chosen steered answer per question still flags 6/6 students under both rules, so the contrast with token watermarks does not depend on selection, whereas selection applied to unsteered answers flags 0/6 and steering along a random direction of the same norm flags 0/3 Llama students (Appendix D.2). At half strength, $\lambda = 0.15$, which raises perplexity by at most 2.8% (Appendix E), the watermark is still detected in 5/6 students, and in 4/6 under the stricter rule.

Imitation Because verification is keyless, a teacher instructed to express e also produces flagged students, but at a perplexity cost of at least 31% against at most 8.1% for our watermark, and once diluted to a cost matching ours† the imitated shift is flagged by neither of the two classifiers that did not select the training data (Appendix C.4).

Erasure Continued fine-tuning on clean data is the strongest attack we tested, leaving 1/6 students detected after 0.5M clean tokens at the cost of moving the student away from the teacher's answers, with perplexity up by 37% to 48%, whereas KGW is already undetected after 50k tokens (Figure 5b and Appendices C.5 and E). When watermarked answers make up only 10% to 50% of the training tokens†, the watermark is still detected in 1/3 to 3/3 Llama students and KGW in none (Figure 5c and Appendix C.6), and token watermarks remain readable from next-token probabilities where answers no longer reveal them, so the two carriers are complementary (Appendix C.7).

6 Conclusion

We pioneered EmotionMark, an emotion-circuit watermark for verifying whether a student was distilled from a watermarked teacher. Steering the teacher's emotion circuit implants a weak affective tendency that students learn from text alone, candidate selection amplifies it, and public emotion classifiers read it from a suspect's ordinary answers. With answer-only verification, it remains detectable in 100% of students after paraphrasing of the distillation data, whereas KGW falls from 100% to at most 50%. It survives output rewriting, and an adaptive attacker guided by the sum of all three classifier scores leaves it in 67% of students while the corresponding attack removes KGW entirely. No false positives occur among 118 newly trained unwatermarked students or 30 students of independent teachers, and perplexity rises by at most 8.1% on the primary training questions. Since token watermarks remain readable from next-token probabilities, the two carriers are complementary. Keying the affective carrier to a private target emotion is a natural next step.

AI USE STATEMENT

We used large language model assistants to help edit the text of this paper. All reported numbers are computed by our code from logged model outputs, and the authors checked every result and every claim in the paper. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

Our watermark lets a model owner detect unauthorized distillation after the fact. It does not restrict how the teacher or its students can be used, and it changes the teacher's answers only by a weak affective shift whose cost we measure (Section 5.3). Because verification is keyless, a positive decision indicates inheritance of the affective shift and serves as grounds for further investigation rather than as proof of distillation by itself (Section 5.4). The WildChat questions we use as distillation prompts come from public user conversations and may contain offensive content.

REPRODUCIBILITY STATEMENT

Section 4 and Appendices A.2, A.3, A.4, and B.1 give the models, prompts, hyperparameters, and decision rule, and Appendix F lists every implementation correction. Retraining seven students with their recorded seeds reproduced their adapter weights byte for byte. We will release the code, the circuit, the seeds, and the per-student results.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Thibaud Ardoin, Jonas Schäfer, and Gerhard Wunder. Llm self-recognition: Steering and retrieving activation signatures. *arXiv preprint arXiv:2606.06315*, 2026.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Roger L Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. Postmark: A robust blackbox watermark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8969–8987, 2024.

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.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Jiayu Ding, Lei Cui, Li Dong, Nanning Zheng, and Furu Wei. Information-preserving reformulation of reasoning traces for antidistillation. *arXiv preprint arXiv:2510.11545*, 2025.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. In *2023 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2023.

Chenchen Gu, XIANG LI, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. In *International Conference on Learning Representations*, volume 2024, pp. 38745–38768, 2024.

Jochen Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. 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. 4067–4082, 2024.

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.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 49660–49704, 2024.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in neural information processing systems*, 36:27469–27500, 2023.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15617–15639, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 6499–6519, 2024.

Jiacheng Lu, Yiming Li, Tao Song, Weijian Wang, Wenjie Qu, Haibing Guan, and Jiaheng Zhang. Echoes within the reasoning: Stealthy and effective watermarking via chain of thought. *arXiv preprint arXiv:2605.28890*, 2026.

Xinhang Ma, William Yeoh, Ning Zhang, and Yevgeniy Vorobeychik. Protecting language models against unauthorized distillation through trace rewriting. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11307–11324, 2026.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. Can llm watermarks robustly prevent unauthorized knowledge distillation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13228–13251, 2025.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *arXiv preprint arXiv:1603.05766*, 2016.

Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 613–625, 2024.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37: 21079–21113, 2024.

Tom Sander, Hongyan Chang, Tomáš Souček, Tuan Tran, Valeriu Lacatusu, Sylvestre-Alvise Rebuffi, Alexandre Mourachko, Surya Parimi, Christophe Ropers, Rashel Moritz, et al. Textseal: A localized llm watermark for provenance & distillation protection. *arXiv preprint arXiv:2605.12456*, 2026.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, et al. Do llms" feel"? emotion circuits discovery and control. *arXiv preprint arXiv:2510.11328*, 2025.

Yixuan Even Xu, John Kirchenbauer, Yash Savani, Asher Trockman, Alexander Robey, Tom Goldstein, Fei Fang, and J Zico Kolter. Antidistillation fingerprinting. *arXiv preprint arXiv:2602.03812*, 2026.

Guang Yang, Amir Ghasemian, Fengchen Liu, Zhong Wang, Ninareh Mehrabi, and Homa Hosseinmardi. Asking back: Interaction-layer antidistillation watermarks. *arXiv preprint arXiv:2605.16462*, 2026.

Zhuohao Yu, Xingru Jiang, Weizheng Gu, Yidong Wang, Qingsong Wen, Shikun Zhang, and Wei Ye. Saemark: Steering personalized multilingual llm watermarks with sparse autoencoders. *Advances in Neural Information Processing Systems*, 38:158702–158731, 2026.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix

This appendix gives the settings, per-student results, and additional analyses behind Section 5. Unless a table states otherwise, every student is verified with the same 400 questions, the same 20 reference students of its family, and the rule $\mathcal{V}$. Results marked with †, and all results under the stricter rule, come from additional analyses defined after the main results were read. The three students of a configuration in one family are called training runs 1, 2, and 3.

A.1 Algorithm

Algorithm 1 gives the generation procedure W and the verification rule $\mathcal{V}$ of Section 4, with e the target emotion and $K = 5$, $N = 400$, and $B = 20$ in our experiments. As in the released code of Wang et al. (2025), the attention term is added at every prompt and answer position, and the MLP term only at positions that predict the next token.

Algorithm 1 Watermarked generation W and answer-only verification $\mathcal{V}$

procedure GENERATE(T, q) ▷ run by the owner for each training question
  for $k = 1, \ldots, K$ **do**
    sample y_k from T on q, adding $\lambda\, m_\ell \odot \delta_\ell$ to $h_{\ell,t}$ in every module at the positions of Section 4.1
  end for
  return the first y_k with the highest $f_{\text{sel}}(y_k)[e]$ ▷ selection (Section 4.1), and $K = 1$, like a random choice among the K samples, is steering alone
end procedure

procedure VERIFY($\tilde{S}$) ▷ $\tilde{S}$ is the suspect service
  $\hat{y}_j \leftarrow$ answer of $\tilde{S}$ to v_j, for $j = 1, \ldots, N$ ▷ reference statistics are computed once in the same way
  for $r = 1, 2, 3$ **do**
    compute d_r and c_r from $\hat{y}_1, \ldots, \hat{y}_N$
    $\text{pos}_r \leftarrow \mathbf{1}\big[\, d_r > \max_{1 \le b \le B} d_{r,b}^{(0)} \ \wedge \ c_r > \max_{1 \le b \le B} c_{r,b}^{(0)} \,\big]$ ▷ same as $p_r \le 0.05$ for $B = 20$
  end for
  return flag if $\text{pos}_1 + \text{pos}_2 + \text{pos}_3 \ge 2$, and no flag otherwise
end procedure

A.2 EMOTION CLASSIFIERS

Table 2 lists the three classifiers. The language-model classifier receives the fixed prompt below inside its chat template and keeps the first and last 768 tokens of any answer longer than 1,536 tokens. The two fine-tuned classifiers read at most 512 tokens, so a longer answer is split into consecutive windows whose probabilities are averaged with weights proportional to window length. Verification answers are generated with at most 256 new tokens, so an answer may stop before the model would have ended it, and such answers are scored as returned and kept in all averages. The language-model classifier is Qwen2.5-7B-Instruct, an open instruction-tuned model with seven billion parameters. The selection classifier is a DistilRoBERTa model, a compact version of the RoBERTa text encoder, fine-tuned to assign each text a probability for each of seven classes, namely anger, disgust, fear, joy, neutral, sadness, and surprise. The GoEmotions classifier is trained on Reddit comments labeled with 27 emotion categories and a neutral class. A whitespace-only answer is replaced by a one-character placeholder for the language-model classifier and is skipped by the other two. Every verification answer of the reported students receives a valid score from all three classifiers. The configuration we will release records the model revisions, the GoEmotions labels used for e and $\bar{e}$, and the tokenizer settings.

Table 2: Emotion classifiers used for selection and verification.

Classifier	Model	Score used
Language-model classifier	`Qwen/Qwen2.5-7B-Instruct`	probability of each option letter, renormalized over the seven letters
Selection classifier	`j-hartmann/emotion-english-distilroberta-base`	probability of each of seven classes
GoEmotions classifier	`SamLowe/roberta-base-go_emotions`	independent probability of each of 28 labels

System. You are a precise emotion classifier.
User. Read the passage below and decide which ONE emotion it most strongly expresses.
Choose exactly one option:
A. anger
B. disgust
C. fear
D. happiness
E. sadness
F. surprise
G. neutral
Passage:
"""
{passage}
"""
Answer with a single letter (A-G):

A.3 TOKEN-LEVEL BASELINES

Each baseline watermarks the teacher's answers to the same 4,096 training questions while sampling, with temperature 0.7, nucleus threshold 0.95, at most 512 new tokens, and one answer per question without selection. KGW follows the reference implementation of Kirchenbauer et al. (2023). SynthID-Text, TextSeal, and Gumbel-max use the implementations released with TextSeal (Sander et al., 2026), and Gumbel-max follows the scheme and score of Fernandez et al. (2023). These three methods share one key and one hash over the three preceding tokens, the setting fixed for this study. ADFP (Xu et al., 2026) is detected with the key used for generation. Table 3 lists the settings.

A detector scores the student's answers to the 400 verification questions pooled together. Scoring starts after the first full context, each pair of context and next token is scored once, and the student is ranked by the method's z-statistic. It is flagged when this statistic exceeds that of all 20 reference students of its family. Gumbel-max sampling is deterministic given the key and the context, so its three training runs share the same teacher answers and differ only in the student's training seed.

Table 3: Token-level baselines. The teacher statistic is computed on 400 watermarked teacher answers of training run 1, the number of answers scored for each student. The student range covers the six students without attack, and the last column gives the largest statistic among the 20 reference students of the Llama and of the Qwen family.

Method	Settings	Teacher	Students	Reference maximum
KGW	$\gamma = 0.5$, $\delta = 4$, context of 2 tokens	138	2.19–4.68	2.03 / 2.49
SynthID-Text	tournament depth 10, context of 3 tokens	44.8	0.12–1.90	2.28 / 2.32
TextSeal	mixing weight 0.1, context of 3 tokens	54.3	-0.29–1.64	1.62 / 1.47
Gumbel-max	context of 3 tokens	60.0	-1.15–0.83	0.97 / 0.94
ADFP	strength 64	83.9	0.78–3.60	1.55 / 1.15

A.4 ATTACKS

Paraphrasing and rewriting. Both rewriters receive the system message "You are a careful editor who rewrites text without changing its meaning." and the instruction below, where the bracketed sentence appears only for the neutral rewriter.

Rewrite the following answer using different wording and sentence structure.
Preserve every fact, number, negation, step of reasoning and the conclusion.
[Remove any emotional or evaluative tone and keep the text objective and neutral.]
Do not add or remove information. Output only the rewritten answer.
Question:
{prompt}
Answer to rewrite:
{answer}

Both use Qwen2.5-7B-Instruct with temperature 0.7, nucleus threshold 0.95, top-k 20, repetition penalty 1.05, and at most 512 new tokens, and the same rewriters rewrite the deployed student's answers. DIPPER (Krishna et al., 2023) uses the released `kalpeshk2011/dipper-paraphraser-xxl` checkpoint with lexical and order diversity 60, temperature 1.0, and nucleus threshold 0.75, and paraphrases long inputs in sentence-aligned chunks of at most 512 tokens that are concatenated in order.

Deployment changes. The two system prompts are "You are a helpful assistant. Answer in a strictly neutral, factual tone and do not express any emotion." and "You are a helpful assistant. Always answer in a warm, cheerful and upbeat tone." Greedy decoding replaces sampling, and the temperature attack samples at temperature 1.0 without nucleus truncation. All other settings, including the limit of 256 new tokens, are unchanged. Logit-level neutralization follows Pan et al. (2025). It compares the next-token distributions of the distilled student and its initial checkpoint on the student's training data, estimates which tokens distillation made more likely after each context of three tokens, and subtracts 2.5 times this estimate from the student's logits at every decoding step, for the verification answers and the benchmarks alike.

Best-of-k rewriting. For each verification question the neutral rewriter produces eight rewrites of the student's answer, and with a budget of k the attacker keeps the rewrite with the least evidence among the first k. When the attacker is guided by the selection classifier, the evidence is its score of e. For the joint attacks, each classifier scores a rewrite by the smaller of its standardized score of e and its standardized contrast between e and $\bar{e}$, using the per-answer mean and standard deviation of five unwatermarked development students of the same family, one per answer source, which are not reference students, and the attacker minimizes the largest or the sum of the three scores. For a token baseline, the evidence is the statistic of its own detector with the correct key. Table 13 gives detection at every budget.

B Reference students and false positives

B.1 Students and reference students

All students use LoRA adapters of rank 128 with $\alpha = 128$ and no dropout on every linear layer except the output head. They are trained with AdamW at learning rate 2×10^{-4}, betas $(0.9, 0.999)$, no weight decay, 3% warmup, and cosine decay, in bfloat16 with sequence length 1024 and effective batch size 32, with the loss on answer tokens only, until two million answer tokens have been consumed. The teacher samples at temperature 0.7 with nucleus threshold 0.95 and at most 512 new tokens, and students answer the verification questions with the same sampling and at most 256 new tokens. The three training runs of a configuration use different sampling and training seeds and share the training questions, the verification questions, and the reference students.

Each reference student draws one of the five answer sources of Section 4.2 uniformly and independently. Besides single unsteered answers of our teacher, two of the sources are answers of Qwen2.5-7B-Instruct and of Llama-3.1-8B-Instruct, an open instruction-tuned model with eight billion parameters, and the remaining two come with public datasets. These are the original answers of Alpaca (Taori et al., 2023), a set of instructions answered by an OpenAI model, from which our primary training questions are drawn, and the assistant answers in WildChat (Zhao et al., 2024), a collection of user conversations with ChatGPT, which supplies our second set of training questions. Each reference student has its own sampling and training seeds. Drawing the source at random makes the reference students vary across these five sources rather than follow a single recipe. Qwen2.5-7B-Instruct plays several roles in this study. It is one of the classifiers, supplies one reference source, serves as the instruction-tuned rewriter in our paraphrasing attacks, and, together with Llama-3.1-8B-Instruct, serves as an independent teacher in Section 5. Table 4 gives the realized draw and Table 5 the realized exposure. Because the budget is fixed in answer tokens, students whose sources give shorter answers take more passes over their 4,096 examples. The 59 newly trained unwatermarked students[†] per family are built in the same way and never used as references, and 20 earlier unwatermarked students per family, trained with the same recipe, serve as further controls. The Qwen tokenizer cannot reproduce the boundary between question and answer in 246 of 1,916,232 training rows (0.013%), all in unwatermarked Qwen students, and these rows

are skipped. No other student skips a row. Appendix B.4 repeats the main decisions with the 40 reference students of both families pooled.

Table 4: Realized answer sources of the 20 reference students of each family.

Students	Unsteered teacher	Qwen2.5-7B	Llama-3.1-8B	Alpaca	WildChat	Total
Llama	2	5	3	1	9	20
Qwen	2	4	6	3	5	20

Table 5: Realized training exposure under the budget of two million answer tokens. Passes are examples seen divided by the size of the sampled training set.

Family	Students	n	Optimizer steps	Passes
Llama	Watermarked	3	283–284	2.210–2.215
Llama	Unsteered teacher	3	284–285	2.215–2.226
Llama	Reference students	20	197–521	1.539–4.065
Llama	Newly trained unwatermarked†	59	196–531	1.531–4.147
Qwen	Watermarked	3	282–283	2.203–2.207
Qwen	Unsteered teacher	3	283–284	2.207–2.217
Qwen	Reference students	20	198–543	1.551–4.236
Qwen	Newly trained unwatermarked†	59	196–536	1.528–4.181

B.2 Decision rule

With $B = 20$, classifier r is positive exactly when d_r and c_r both exceed the corresponding statistics of all 20 reference students (Section 4.2). We use $B = 20$, which places the smallest attainable p-value just below the threshold, because each reference student costs one fine-tuning run. Suppose the suspect is exchangeable with the reference students, meaning that it could have been produced by the same randomized recipe, so that every ordering of the $B + 1$ students is equally likely. Then the suspect exceeds all B reference values of a statistic with probability at most $1/(B+1)$ (Phipson & Smyth, 2016). A classifier is positive only when both of its comparisons succeed, so taking the larger p-value is an intersection-union test, which keeps this bound without a correction for using two statistics (Berger, 1982). The three classifiers score the same answers, so their decisions are dependent, and we neither combine their p-values nor assume independence. Let X be the number of positive classifiers. Under any dependence $\mathbb{E}[X] \le 3/(B+1) = 3/21$, and Markov's inequality gives $\Pr[X \ge 2] \le \mathbb{E}[X]/2 = 3/(2(B+1)) = 1/14 \approx 0.071$. If the classifiers erred independently, $\Pr[X \ge 2] \le 3/(B+1)^2 = 3/21^2 \approx 0.007$. Both bounds average over the random draw of the reference students, whereas a verifier uses one fixed set, whose false-positive rate can differ from this average (Bates et al., 2023). Tables 6 to 8 therefore measure the false-positive rate of the fixed reference students directly. The estimate on the 59 newly trained unwatermarked students† per family refers to the same mixture of five sources, and Table 7 also reports it for each source. The fixed set does matter for a single classifier, so for every group of unwatermarked students we report both the decision of $\mathcal{V}$ and the number of students that at least one classifier finds positive. No newly trained unwatermarked student† is flagged, which gives a one-sided 95% Clopper–Pearson upper bound of 4.95% on the false-positive rate for the 59 students of one family and 2.51% for the 118 students of both.

B.3 Unwatermarked students

Table 6 gives unwatermarked students without attack and Table 7 the newly trained ones by answer source. Of the 30 students of independent teachers, 12 were trained on the teachers' ordinary answers, generated in the same way as two of the reference sources, 12 on their answers to questions on negative topics, and 6 on answers of the Qwen teacher selected for e without steering. No classifier is positive for any of the 59 newly trained Qwen students. The selection classifier alone is positive for 9 of the 59 newly trained Llama students, all trained on the original Alpaca answers, a source that supplies one of the 20 Llama reference students, and $\mathcal{V}$ flags none of them. Table 8 gives unwatermarked students subjected to each attack. Apart from DIPPER, two are flagged, one of 20

Qwen students after continued fine-tuning on two million clean tokens, which remains flagged under the stricter rule and the pooled reference students, and one of 20 Llama students after pruning half of its weights, which neither the stricter rule nor the pooled reference students flag. After DIPPER, one earlier Llama student and one of the six students trained on unsteered answers of our teacher are flagged, and Table 12 recalibrates this condition.

Table 6: Unwatermarked students without attack. Columns give the students flagged under our rule and under the stricter rule, the students positive under each classifier, and those positive under at least one. LM denotes the language-model classifier.

Students	n	Our rule	Stricter	LM	Selection	GoEmotions	Any
Newly trained, Llama†	59	0	0	0	9	0	9
Newly trained, Qwen†	59	0	0	0	0	0	0
Earlier unwatermarked, Llama	20	0	0	0	2	0	2
Earlier unwatermarked, Qwen	20	0	0	0	0	1	1
Independent teachers, ordinary answers	12	0	0	0	0	1	1
Independent teachers, negative topics	12	0	0	1	0	0	1
Qwen teacher, selection without steering	6	0	0	0	1	0	1
Unsteered answers of our teacher	6	0	0	0	0	0	0

Table 7: Newly trained unwatermarked students by answer source (additional analysis†). None is flagged under our rule or the stricter rule.

	Llama students		Qwen students	
Answer source	n	Any classifier	n	Any classifier
Unsteered teacher	10	0	8	0
Qwen2.5-7B	14	0	19	0
Llama-3.1-8B	13	0	10	0
Alpaca	11	9	12	0
WildChat	11	0	10	0
Total	59	9	59	0

B.4 Reference students of both families

In an additional analysis†, a classifier is positive when both of its statistics exceed those of all 40 reference students of both families. Pooling can only raise the two thresholds, so it removes flags but never adds them. Table 9 lists every reported condition whose count changes, together with the unwatermarked students and the students of teachers instructed to answer with disgust. Every condition for our watermark in Table 1 is unchanged, except the two best-of-eight attacks guided by all three classifiers, which were not recomputed.

Table 8: Unwatermarked students subjected to each attack and verified against the unchanged reference students. Flagged counts students flagged by $\mathcal{V}$, which are also the counts under the stricter rule except for the pruned Llama student, and Any counts students with at least one positive classifier. For neutral paraphrasing, the 79 students per family are the 59 newly trained and the 20 earlier unwatermarked students, of which Table 1 counts only the newly trained. Model modifications apply each of the five modifications to the 20 earlier unwatermarked students of each family, with 93 of the 100 Qwen evaluations completed. Students trained on unsteered answers of our teacher are counted in Table 1 for the conditions listed there. Of these six students, at least one classifier is positive for three after DIPPER and for three after logit-level neutralization, and for none after any other attack.

	Llama students			Qwen students		
Attack	n	Flagged	Any	n	Flagged	Any
Ordinary paraphrasing	5	0	0	5	0	0
Neutral paraphrasing	79	0	17	79	0	6
DIPPER	20	1	18	20	0	8
Ordinary rewriting	20	0	0	20	0	0
Neutral rewriting	20	0	0	20	0	0
Best-of-eight, selection classifier	20	0	0	20	0	0
Best-of-eight, KGW detector	20	0	1	20	0	0
Best-of-eight, ADFP detector	20	0	1	20	0	0
Neutral system prompt†	5	0	2	5	0	1
Cheerful system prompt†	5	0	1	5	0	1
Greedy decoding†	5	0	2	5	0	0
Temperature 1.0†	5	0	2	5	0	0
Logit-level neutralization	20	0	1	20	0	0
Model modifications†	100	1	9	93	0	2
Continued fine-tuning, 0.5M tokens	20	0	4	20	0	4
Continued fine-tuning, 2M tokens	20	0	2	20	1	2

Table 9: Students flagged by $\mathcal{V}$ with the 20 reference students of the suspect's own family and with the 40 reference students of both families pooled (additional analysis†).

Students	Own family	Both families
Our watermark, changed counts		
Half strength, $\lambda = 0.15$	5/6	4/6
Merge with the initial checkpoint	5/6	4/6
Continued fine-tuning, 0.5M tokens	1/6	0/6
Watermarked share 10%, Llama students	1/3	0/3
Watermarked share 25%, Llama students	2/3	1/3
Unwatermarked and token-watermarked students		
Newly trained unwatermarked	0/118	0/118
Earlier unwatermarked	0/40	0/40
Independent teachers	0/30	0/30
Unwatermarked, DIPPER	2/46	1/46
Token watermarks, DIPPER	3/30	0/30
Unwatermarked Qwen, continued fine-tuning 2M	1/20	1/20
Unwatermarked Llama, pruning 50%	1/20	0/20
Teachers instructed to answer with disgust		
Our teacher and independent teachers, full or milder instruction	30/30	30/30
Milder instruction mixed in at a cost comparable to ours, Llama students	3/6	0/6
Milder instruction mixed in at 50%, Llama teacher, Llama students	3/3	0/3

C ADDITIONAL RESULTS

C.1 PARAPHRASING BEFORE DISTILLATION

Table 10 gives the decision for every student without attack and after the three paraphrasings of Table 1, including SynthID-Text, TextSeal, and Gumbel-max. Table 11 breaks down the teacher-side analysis of Section 5.2 by classifier and by token method, and Table 12 recalibrates the DIPPER students.

Table 10: Decision for every student without attack and after the three paraphrasings of Table 1. Each entry lists training runs 1, 2, and 3 of a family, with 1 for a flagged student.

Method	Paraphrasing	Llama	Qwen	Total
Ours	None	1 / 1 / 1	1 / 1 / 1	6/6
Ours	Ordinary	1 / 1 / 1	1 / 1 / 1	6/6
Ours	Neutral	1 / 1 / 1	1 / 1 / 1	6/6
Ours	DIPPER	1 / 1 / 1	1 / 1 / 1	6/6
KGW	None	1 / 1 / 1	1 / 1 / 1	6/6
KGW	Ordinary	0 / 1 / 0	0 / 0 / 0	1/6
KGW	Neutral	1 / 0 / 1	1 / 0 / 0	3/6
KGW	DIPPER	0 / 0 / 0	0 / 0 / 0	0/6
SynthID-Text	None	0 / 0 / 0	0 / 0 / 0	0/6
SynthID-Text	Ordinary	0 / 0 / 0	0 / 0 / 0	0/6
SynthID-Text	Neutral	0 / 0 / 0	0 / 0 / 0	0/6
SynthID-Text	DIPPER	0 / 0 / 0	0 / 0 / 0	0/6
TextSeal	None	0 / 1 / 0	0 / 0 / 0	1/6
TextSeal	Ordinary	0 / 0 / 0	0 / 0 / 0	0/6
TextSeal	Neutral	0 / 0 / 0	0 / 0 / 0	0/6
TextSeal	DIPPER	0 / 1 / 0	0 / 0 / 0	1/6
Gumbel-max	None	0 / 0 / 0	0 / 0 / 0	0/6
Gumbel-max	Ordinary	0 / 0 / 0	0 / 0 / 0	0/6
Gumbel-max	Neutral	0 / 0 / 1	0 / 0 / 0	1/6
Gumbel-max	DIPPER	0 / 0 / 0	0 / 0 / 1	1/6
ADFP	None	1 / 1 / 1	1 / 1 / 0	5/6
ADFP	Ordinary	0 / 0 / 0	0 / 0 / 0	0/6
ADFP	Neutral	0 / 0 / 0	0 / 0 / 0	0/6
ADFP	DIPPER	0 / 0 / 0	1 / 0 / 0	1/6

Table 11: Share of the teacher-side signal that remains after paraphrasing, on the 4,096 teacher answers of each training run (4,094 to 4,096 after DIPPER, which leaves a few answers empty), averaged over the three runs (additional analysis†). For our watermark, the signal is the difference in the mean score of e between watermarked and unsteered teacher answers, both paraphrased in the same way, under each classifier. For a token watermark, it is the detection statistic divided by the square root of the number of scored tokens, minus the same quantity on unwatermarked answers.

Signal	Ordinary	Neutral	DIPPER
Language-model classifier	66%	55%	85%
Selection classifier	56%	55%	63%
GoEmotions classifier	38%	33%	72%
Our watermark, mean	53%	48%	73%
KGW	18%	25%	8%
SynthID-Text	15%	23%	5%
TextSeal	15%	23%	6%
Gumbel-max	14%	23%	5%
Token watermarks, mean	16%	24%	6%

DIPPER raises the emotion scores of every student trained on its output. In an additional analysis†, each DIPPER student is therefore compared with the 20 earlier unwatermarked students of its family trained on DIPPER paraphrases of their own training answers, instead of the unchanged reference students.

Table 12: Students trained on DIPPER paraphrases, flagged by our rule and by the stricter rule against the unchanged reference students and against 20 unwatermarked DIPPER students per family (additional analysis†). Token-watermarked rows apply $\mathcal{V}$, not the token detector.

	Unchanged reference students		DIPPER reference students	
Students	Our rule	Stricter	Our rule	Stricter
Ours	6/6	6/6	6/6	6/6
Unsteered answers of our teacher	1/6	1/6	1/6	1/6
KGW	0/6	0/6	0/6	0/6
SynthID-Text	0/6	0/6	0/6	0/6
TextSeal	1/6	0/6	0/6	0/6
Gumbel-max	1/6	0/6	0/6	0/6
ADFP	1/6	0/6	0/6	0/6

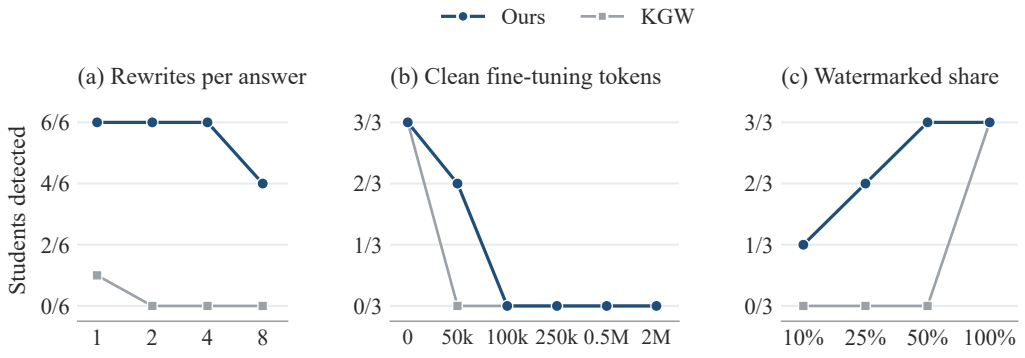


Figure 5: Detection as the attacker's budget grows, for our watermark and for KGW, the only token baseline detected in all six students without attack. (a) Best-of-k rewriting of the deployed student's answers, six students, with our watermark attacked under the strongest guidance we tested, the sum of all three classifier scores†, and KGW under its own detector. (b) Continued fine-tuning on clean data, three Llama students, with budgets below 0.5M tokens from an additional analysis†. (c) Share of watermarked answers in the training data, three Llama students†. Counts under the stricter rule, for the other guidance modes, and for the remaining token baselines are in Tables 13, 17, and 18.

C.2 Attacks on the student

Figure 5a and Table 13 expand the best-of-eight rows of Table 1 over the candidate budget. Under our rule, our watermark remains detected at every budget when the attacker is guided by the selection classifier or by the largest of the three scores, and guidance by the sum of the three scores lowers detection only at eight candidates per answer. Under the stricter rule, detection drops at eight candidates for the largest score and from four candidates for the sum. Under guidance by the selection classifier, the other two classifiers remain positive for every one of our students. No unwatermarked student attacked with guidance from the selection classifier is flagged at any budget.

Table 13: Best-of-k rewriting of the deployed student's answers by the neutral rewriter. With $k = 1$ the first of the eight rewrites is kept without selection. For our watermark each cell gives our rule and the stricter rule, and for a token baseline its own detector, which also guides the selection.

Guidance	$k = 1$	$k = 2$	$k = 4$	$k = 8$
Ours, selection classifier	6/6, 6/6	6/6, 6/6	6/6, 6/6	6/6, 6/6
Ours, largest of three classifiers†	6/6, 6/6	6/6, 6/6	6/6, 6/6	6/6, 3/6
Ours, sum of three classifiers†	6/6, 6/6	6/6, 6/6	6/6, 5/6	4/6, 2/6
KGW, own detector	1/6	0/6	0/6	0/6
ADFP, own detector	1/6	1/6	1/6	0/6

C.3 Model modifications

Each of the six students without attack is quantized to 8 bits or to 4-bit NF4, pruned with Wanda at 20% or 50% sparsity, or merged with its initial checkpoint at equal weights, giving 30 evaluations of the same six students (Table 14). Applied to the 20 earlier unwatermarked students of each family, the same modifications† flag one of 193 evaluations, a Llama student pruned at 50%, and they flag none of the 30 evaluations of students trained on unsteered answers. Pruning half of the weights raises the perplexity of the Llama students' answers by 34% to 42% relative to the same students before pruning (Table 25) and leaves almost none of their ARC-Challenge answers in a parsable format, yet all three remain detected, as do the three Qwen students pruned at 50%.

Table 14: Detection after modifying the weights of the six students without attack. For our watermark each cell gives our rule and the stricter rule, and KGW pools both families.

Modification	Llama	Qwen†	KGW
8-bit quantization	3/3, 3/3	3/3, 3/3	6/6
4-bit quantization (NF4)	3/3, 3/3	3/3, 3/3	6/6
Pruning, 20%	3/3, 3/3	3/3, 3/3	6/6
Pruning, 50%	3/3, 3/3	3/3, 3/3	4/6
Merge with the initial checkpoint	3/3, 2/3	2/3, 2/3	1/6
Total	15/15, 14/15	14/15, 14/15	23/30

C.4 Filtering and imitation

Filtering. The distiller discards the half of the teacher's answers with the highest score of e under the selection classifier and trains on the rest. All six students remain flagged under both rules, students trained on unsteered answers filtered in the same way are flagged in none of six cases, and the perplexity of the watermarked students' answers is 5.0% to 5.9% higher than that of these students. Because the budget is fixed, filtered students pass over their halved training set about 4.6 times.

Imitation. Table 15 lists students of teachers instructed to answer with disgust. Their perplexity rises by 31% to 80%, and their answers lose more distinct bigrams and repeat more trigrams than those of our students. Table 16 mixes answers of the milder instruction into the same teacher's ordinary answers. At the two shares whose cost is comparable to ours, 25% for the Llama teacher and 10% for the Qwen teacher, three of six students are flagged under our rule. None is flagged under the stricter rule or against the pooled reference students (Table 9), and each scores lower under the language-model classifier than every Llama student with our watermark.

Table 15: Students of our watermarked teacher and of teachers instructed to answer with disgust, three per family and teacher. Perplexity is relative to students trained on unsteered answers of our teacher. The last two columns give the largest drop in the share of distinct bigrams and the largest rise in the share of repeated trigrams, whose margins are 0.02.

Teacher	Instruction	n	Our rule	Stricter	Perplexity ratio	Distinct bigrams	Repeated trigrams
Our teacher, steered (ours)	none	6	6/6	6/6	1.06–1.08	0.004	0.007
Our teacher	full	6	6/6	6/6	1.45–1.49	0.152	0.127
Llama-3.1-8B	full	6	6/6	6/6	1.46–1.48	0.130	0.106
Qwen2.5-7B	full	6	6/6	6/6	1.68–1.74	0.119	0.091
Llama-3.1-8B	milder	6	6/6	6/6	1.31–1.32	0.066	0.041
Qwen2.5-7B	milder	6	6/6	5/6	1.70–1.80	0.026	0.016

C.5 Continued fine-tuning

The attacker fine-tunes a distilled student on clean instruction data from databricks-dolly-15k with the same recipe, each budget in a separate run from the distilled student, and the reference students and threshold stay unchanged (Figure 5b and Table 17). Across both families, one of six students remains detected after 0.5M and after 2M clean tokens, in both cases a Qwen student.

Table 16: Llama students when answers of the milder instruction make up the given share of the training tokens and the rest are the same teacher's ordinary answers (additional analysis†). Perplexity is relative to students trained on unsteered answers of our teacher, and the last column is the mean score of e under the language-model classifier (LM).

Teacher	Share	Our rule	Stricter	Perplexity ratio	LM score
Llama-3.1-8B	10%	0/3	0/3	0.99–1.01	0.024–0.044
Llama-3.1-8B	25%	1/3	0/3	1.03–1.04	0.061–0.074
Llama-3.1-8B	50%	3/3	0/3	1.11	0.119–0.124
Llama-3.1-8B	100%	3/3	3/3	1.31	0.236–0.265
Qwen2.5-7B	10%	2/3	0/3	1.08–1.09	0.040–0.059
Qwen2.5-7B	25%	3/3	0/3	1.15–1.16	0.091–0.112
Qwen2.5-7B	50%	3/3	1/3	1.26–1.29	0.186–0.228
Qwen2.5-7B	100%	3/3	2/3	1.76–1.80	0.385–0.390
Ours	–	3/3	3/3	1.06–1.08	0.114–0.152

Table 17: Detection after continued fine-tuning on clean data. KGW is tested with its own detector. The last column counts unwatermarked students fine-tuned in the same way and flagged by $\mathcal{V}$.

Students	Clean tokens	Ours	Ours, stricter	KGW	Unwatermarked
Llama	0	3/3	3/3	3/3	–
Llama	50k†	2/3	0/3	0/3	0/5
Llama	100k†	0/3	0/3	0/3	0/5
Llama	250k†	0/3	0/3	0/3	0/5
Llama	0.5M	0/3	0/3	0/3	0/20
Llama	2M	0/3	0/3	0/3	0/20
Both families	0.5M	1/6	0/6	0/6	0/40
Both families	2M	1/6	1/6	0/6	1/40

C.6 Dilution

In an additional analysis†, watermarked answers are mixed with unsteered answers of the same teacher so that they make up 10%, 25%, or 50% of the two million training tokens of Llama students, and the token baselines are mixed in the same way. Because the budget is fixed, a lower share also means fewer distinct watermarked questions and fewer passes over them, so the three shares form a nested curve rather than varying the proportion alone (Figure 5c and Table 18).

Table 18: Detection of Llama students when watermarked answers form part of the training tokens, three students per cell. Token baselines are tested with their own detectors. Question counts give the range over training runs, mean passes divide the watermarked examples seen by the number of watermarked questions, and perplexity of our students is relative to students trained only on unsteered answers.

Share	Watermarked questions	Mean passes	Perplexity ratio	Ours	Ours, stricter	KGW	SynthID-Text	TextSeal	Gumbel-max
10%	858–948	1.00	1.00–1.01	1/3	0/3	0/3	0/3	0/3	0/3
25%	2,253–2,301	1.00	1.01	2/3	1/3	0/3	0/3	0/3	0/3
50%	4,096	1.11	1.01–1.02	3/3	3/3	0/3	1/3	0/3	1/3
100%	4,096	2.21	1.06–1.08	3/3	3/3	3/3	0/3	1/3	0/3

C.7 Richer access to the student

In an additional analysis†, the owner feeds a Llama student prefixes of newly generated watermarked teacher answers to the 4,096 training questions. The student either samples its next token or exposes its next-token logits, from which the most probable token is taken, and the resulting tokens are scored with each token method's own detector. For Gumbel-max, whose sampling is deterministic, these prefixes coincide with the training answers. Each student is flagged when its statistic exceeds that of all unwatermarked students given the same treatment, 20 per attack and the 40 reference and earlier students without attack, because attacked unwatermarked students already score higher at these interfaces. Table 19 reports 100,000 scored positions. Over the full recorded trace of 0.71 to

1.43 million positions, every cell is 3/3. Students trained on unsteered answers and given the same treatment are flagged in 6 of 120 comparisons (5%) at 100,000 positions and in 10 of 120 (8%) over the full trace. Ordinary paraphrasing is omitted because only five unwatermarked students per family received it, too few for any student to reach significance. ADFP and the Qwen students were not evaluated at these interfaces.

Table 19: Detection with richer access to three Llama students per cell at 100,000 scored positions (additional analysis†). Each cell gives the count with a sampled next token and with the most probable token under the exposed logits.

Condition	KGW	SynthID-Text	TextSeal	Gumbel-max
No attack	3/3, 3/3	3/3, 3/3	3/3, 3/3	3/3, 3/3
Neutral paraphrasing	3/3, 3/3	3/3, 3/3	3/3, 3/3	3/3, 3/3
DIPPER	0/3, 3/3	1/3, 2/3	3/3, 1/3	3/3, 3/3
Continued fine-tuning, 0.5M	3/3, 3/3	3/3, 3/3	3/3, 3/3	3/3, 3/3
Continued fine-tuning, 2M	3/3, 3/3	3/3, 3/3	3/3, 3/3	3/3, 3/3

C.8 Number of verification answers

Table 20 reports how many answers a Llama student needs before detection becomes stable. For $N \in \{50, 100, 200, 400, 800\}$ we apply the decision rule to the answers to the first N of 800 held-out questions, whose first 400 are the verification questions, and record the smallest N from which the student is flagged at every larger budget while the matching student trained on unsteered answers is not. Our watermark reaches this point at the smallest budget, 50 answers, in all three training runs, and needs at most 100 after neutral paraphrasing. With 100 random subsets of an 800-question pool per budget instead, the ratio of the smallest stable budget after neutral paraphrasing to that without attack is 1 for our watermark in every run when detection is required in 50% or 80% of the subsets, and 1, 2, and 2 at 90%, whereas this ratio for KGW at 50% is 2, at least 16, and 8.

Table 20: Smallest number of answers from which Llama students are stably detected, without attack and after neutral paraphrasing of the teacher's answers. The ratio divides the second value by the first, with a paired bootstrap 95% interval over questions for our watermark. A value above 800 lies beyond the largest budget.

	Ours			KGW		
Training run	None	Neutral	Ratio	None	Neutral	Ratio
1	50	100	2.0 $[1, 4]$	400	400	1.0
2	50	50	1.0 $[1, 2]$	50	> 800	≥ 16
3	50	50	1.0 $[0.25, 4]$	100	> 800	≥ 8

For KGW in the second run, the paraphrased student is not flagged at 400 or 800 answers. In the third run it is flagged from 100 to 400 answers, and the value is censored because the matching unwatermarked student is also flagged at 800 answers. With the 20 earlier unwatermarked students as reference students, the KGW ratios are 1, at least 4, and 2. When the paraphrased students are compared with these students after the same paraphrasing, the KGW ratios are at least 2, at least 4, and at least 8, against 4, 1, and 1 for our watermark. At 400 answers, our watermark is detected in all three runs under all three sets of reference students, whereas KGW is detected in two, three, and one of them.

C.9 Verification questions by type

Five human reviewers independently checked whether each verification question is eligible, and their decisions agreed. In an additional analysis†, Qwen2.5-7B-Instruct labels each verification question as having a fixed answer, as open, or as incomplete, giving 131, 239, and 30 questions. When the watermarked and reference students are scored on the questions of one type only, our watermark is detected in all six students for every type, and none of the 40 earlier unwatermarked students is flagged.

D DESIGN CHECKS

D.1 CANDIDATE SELECTION

Table 21 separates steering from selection. Steering followed by a randomly chosen candidate flags all six students, whereas selection applied to unsteered answers flags none. After neutral paraphrasing of the teacher's answers, $\mathcal{V}$ flags both variants, and the stricter rule flags three of three students trained with selection and one of three trained without it, so selection strengthens the signal that steering creates once the teacher's answers are paraphrased. Because the selection classifier is also one of the three verification classifiers, a positive decision could partly reflect the preferences of this single classifier rather than a broader shift in tone. We check this with the stricter rule, which ignores the selection classifier and requires the other two classifiers to be positive, and with a different selector. When the GoEmotions classifier selects the candidates instead, the two selectors pick the same candidate for 39.8% to 40.7% of the training questions in the three runs, against 20% expected for two independent choices among five candidates, and all six students are flagged, each by the language-model classifier and by the classifier of Hartmann (2022), neither of which took part in selection in this check.

Table 21: Steering and selection. The last column requires the two classifiers that play no part in selection to be positive, which is the stricter rule except in the row with GoEmotions selection.

Generation	Students	Our rule	Non-selecting classifiers
Steering and selection (ours)	6	6/6	6/6
Selection without steering	6	0/6	0/6
Steering, random candidate	6	6/6	6/6
Neither steering nor selection	3 Llama	0/3	0/3
Steering, GoEmotions selection†	6	6/6	6/6
Half strength, $\lambda = 0.15$	6	5/6	4/6
Neutral paraphrasing, with selection†	3 Llama	3/3	3/3
Neutral paraphrasing, random candidate†	3 Llama	3/3	1/3

D.2 EMOTION CIRCUIT AND COMPONENT ABLATIONS

The procedure of Wang et al. (2025) estimates an emotion direction in each layer, ranks MLP neurons by their contribution to this direction and attention heads by how much zeroing each head lowers the projection onto it, and distributes a fixed budget of components over the layers according to their influence on the model's final emotion representation. The disgust circuit contains 392 MLP neurons and 168 attention heads. Every layer contributes at least four neurons and two heads, and layers 14 to 27 hold 82% of the neurons and 82% of the heads (Table 22). In an MLP module each coordinate of $h_{\ell,t}$ belongs to one neuron, and in an attention module the coordinates are grouped by head. To compute the difference vectors, the model is instructed through its system prompt to reply to each scenario of the localization dataset of Wang et al. (2025) with each of six emotions in turn. The component list of the circuit has SHA-256 digest

```
79c6a6bb23de5a4433bcb7d52fe39ffbb9f4f3a9282ba876d943eaeebde15b1e
```

Before generation, the loader checks the model geometry, layer coverage, component counts, indices, tensor shapes, and finite nonzero values, and it adds the unnormalized difference vectors at the positions of Section 4.1.

Table 22: Components of the disgust circuit in each layer.

Layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13
MLP neurons	4	4	4	4	4	4	4	5	5	5	5	5	7	10
Attention heads	2	2	2	2	2	2	2	2	2	2	2	2	3	4
Layer	14	15	16	17	18	19	20	21	22	23	24	25	26	27
MLP neurons	13	15	18	20	24	26	27	28	29	28	27	26	22	19
Attention heads	5	6	7	8	9	10	11	12	12	12	12	12	11	10

For Llama students, we also steer with the attention heads or the MLP neurons of the circuit alone, keeping their vectors from the full circuit, or with a random direction of the same norm on the circuit coordinates, each followed by the same selection (Table 23). The teacher-side shift is the mean score of e under the selection classifier over all five candidates minus that of five unsteered candidates. Steering the attention heads alone carries about half of the shift of the full circuit and suffices for detection, whereas the MLP neurons alone and the random direction leave almost no shift and flag no student.

Table 23: Steering parts of the circuit, three Llama students per row. A partial condition removes the other part entirely rather than matching total strength, and the MLP term acts only at positions that predict the next token, so the rows do not compare the two parts at equal dose.

Steered components	Teacher-side shift	Our rule	Stricter
Full circuit	0.104	3/3	3/3
Attention heads only	0.049	3/3	2/3
MLP neurons only	0.006	0/3	0/3
Random direction, same norm	−0.001	0/3	0/3

D.3 Choice of target emotion and strength

We use disgust because unwatermarked answers rarely express it (Table 26), so a small shift can stand out against the reference students. In the circuit-modulation experiments of Wang et al. (2025) on the same teacher, disgust is expressed in all held-out generations in positive, neutral, and negative contexts, as are fear, sadness, and surprise. Another target emotion would need its own circuit, strength, and contrast emotion, and the selection and the reference statistics would use the scores of that emotion.

Choice of strength. The difference vectors are not normalized, so at $\lambda = 0.30$ each circuit coordinate moves by 30% of the average gap between its value under the instruction for e and its mean over the six instructions, which is 30% of the strength $\lambda = 1$ that Wang et al. (2025) use on the same model to make it express an emotion clearly. The strength was chosen before any reported student was trained, on earlier development students distilled with a different training recipe at strengths from 0.15 to 0.90, one student per strength, and on development answers of the teacher. Among the development students, strengths below 0.30 left the scores of e close to those of unwatermarked students, and strengths of 0.45 and above made the answers overtly emotional. For 64 development questions the teacher then sampled five candidates at each of $\lambda \in \{0.15, 0.30, 0.45\}$, scored by the selection classifier, and we read the answers. At $\lambda = 0.15$ the first steered sample scored above the first unsteered sample of the same question on 31 of the 64 questions and below it on 32, at $\lambda = 0.45$ the answers drifted off topic and inserted expressions of e into unrelated tasks, and at $\lambda = 0.30$ the score rose on 42 questions without visible damage to the answers. At $\lambda = 0.30$ the mean score of e over all 320 steered candidates was 0.129 against 0.045 for the unsteered candidates, while the median moved only from 0.014 to 0.020, so the shift sits in the mean and in a minority of answers rather than in the typical answer. We therefore use $\lambda = 0.30$ as the main strength and $\lambda = 0.15$ as the secondary one.

E Quality

Perplexity is computed with Phi-3-mini-4k-instruct on each student's answers to the 800 held-out questions, whose first 400 are the verification questions, and divided by that of the student trained on unsteered answers of the same teacher in the same training run and given the same questions, attack, or modification. A student is within the margins fixed in advance when this ratio is at most 1.10, the share of distinct bigrams drops by at most 0.02, and the share of repeated trigrams rises by at most 0.02. Every unattacked watermarked student is within all three margins, with perplexity ratios of 1.056 to 1.081 on the primary training questions and 1.055 to 1.094 on the WildChat questions (Table 25).

ARC-Challenge and TruthfulQA in the MC1 format are compared with the same unsteered student, with a paired bootstrap interval over questions. A comparison passes when the lower end of the

95% interval for the accuracy ratio is at least 0.95, fails when the upper end is below 0.95, and is inconclusive otherwise (Table 24). For reference, the same rule applied to students trained on unsteered answers in different training runs gives 4 passes, 2 inconclusive, and no failure for Llama on ARC-Challenge, with ratios of 0.97 to 1.03, and 3, 3, and 0 on TruthfulQA. For Qwen it gives 3, 1, and 2 on ARC-Challenge and 4, 2, and 0 on TruthfulQA. An ARC-Challenge answer counts as correct only when it is a single option letter that matches the key. Qwen students often give the letter followed by the option text, which the parser counts as wrong. The failing Qwen student leaves 744 of 1,172 answers unparsed, against 154 for its unsteered counterpart, and 20 of 21 comparisons between other unwatermarked Qwen students and the unsteered student of their training run fail in the same way. ARC-Challenge ratios of Qwen students, including the pass in training run 3, therefore mainly reflect answer format, whereas TruthfulQA is scored by answer likelihood without parsing and gives no failure.

Table 26 gives the share of single answers to which each classifier assigns e a probability above one half. With the language-model classifier, this share averages 13.2% for watermarked students, 4.4% at $\lambda = 0.15$, and 2.3% for students trained on unsteered answers.

Table 24: Accuracy of each watermarked student relative to the student trained on unsteered answers in the same training run, with paired-bootstrap 95% intervals. Students trained on unsteered answers reach 49% to 50% on ARC-Challenge and 23% to 25% on TruthfulQA for Llama, and 54% to 65% and 29% to 30% for Qwen.

Students	Run	Benchmark	Ratio	95% interval	Decision
Llama	1	ARC-Challenge	0.984	[0.940, 1.029]	Inconclusive
Llama	1	TruthfulQA	1.016	[0.951, 1.089]	Pass
Llama	2	ARC-Challenge	0.986	[0.945, 1.029]	Inconclusive
Llama	2	TruthfulQA	1.033	[0.971, 1.099]	Pass
Llama	3	ARC-Challenge	0.951	[0.911, 0.991]	Inconclusive
Llama	3	TruthfulQA	0.935	[0.878, 0.990]	Inconclusive
Qwen	1	ARC-Challenge	0.370	[0.335, 0.405]	Fail
Qwen	1	TruthfulQA	0.953	[0.897, 1.009]	Inconclusive
Qwen	2	ARC-Challenge	0.929	[0.883, 0.977]	Inconclusive
Qwen	2	TruthfulQA	0.957	[0.901, 1.017]	Inconclusive
Qwen	3	ARC-Challenge	1.099	[1.058, 1.142]	Pass
Qwen	3	TruthfulQA	0.953	[0.896, 1.013]	Inconclusive

Five human reviewers also independently assessed paired answers of the students under a fixed rubric, and their decisions agreed. This assessment plays no part in verification.

Table 25: Perplexity ratios of watermarked students. The first ratio is relative to the unwatermarked student given the same questions, attack, or modification, and the second to the same watermarked student without the attack or modification. The last column counts students within all three margins relative to the unwatermarked student given the same questions, attack, or modification.

Condition	Students	vs. unwatermarked	vs. unattacked	Within margins
No attack	Llama	1.056–1.081	–	3/3
No attack	Qwen	1.066–1.073	–	3/3
WildChat questions	Llama	1.055–1.083	–	3/3
WildChat questions	Qwen	1.080–1.094	–	3/3
Half strength, $\lambda = 0.15$	Llama	1.009–1.022	–	3/3
Half strength, $\lambda = 0.15$	Qwen	1.020–1.028	–	3/3
Filtering	Llama	1.053–1.059	–	3/3
Filtering	Qwen	1.050–1.059	–	3/3
Ordinary paraphrasing	Llama	1.036–1.038	1.130–1.166	3/3
Ordinary paraphrasing	Qwen	1.020–1.039	1.128–1.141	3/3
Neutral paraphrasing	Llama	1.037–1.045	1.103–1.146	3/3
Neutral paraphrasing	Qwen	1.021–1.051	1.095–1.109	3/3
DIPPER	Llama	0.981–1.003	1.597–1.668	3/3
DIPPER	Qwen	0.975–1.033	1.550–1.614	3/3
8-bit quantization	Llama	1.056–1.082	0.994–1.012	3/3
4-bit quantization (NF4)	Llama	1.064–1.092	1.050–1.074	3/3
Pruning, 20%	Llama	1.067–1.072	1.005–1.031	3/3
Pruning, 50%	Llama	1.004–1.081	1.345–1.417	0/3
Merge with the initial checkpoint	Llama	1.007–1.030	0.950–0.956	3/3
Continued fine-tuning, 0.5M	Llama	0.998–1.013	1.440–1.479	3/3
Continued fine-tuning, 0.5M	Qwen	0.971–1.008	1.373–1.385	2/3
Continued fine-tuning, 2M	Llama	0.990–1.025	1.976–2.045	3/3
Continued fine-tuning, 2M	Qwen	0.977–1.021	2.082–2.109	3/3

Table 26: Share of verification answers to which a classifier assigns e a probability above one half, as the range over six students. LM denotes the language-model classifier.

Students	LM	Selection	GoEmotions
Ours, $\lambda = 0.30$	11.3–15.5%	15.0–17.8%	0.8–2.0%
Ours, $\lambda = 0.15$	4.0–5.0%	4.8–5.5%	0.0%
Ours, after filtering	6.5–8.3%	7.8–10.0%	0.0–0.5%
Unsteered answers of our teacher	1.8–2.8%	1.0–3.3%	0.0%
Our teacher instructed to answer with disgust	92.5–94.3%	85.0–88.5%	6.8–10.3%

F IMPLEMENTATION CORRECTIONS

The settings of W and $\mathcal{V}$ were fixed before the results for the reported students were examined. Table 27 lists the implementation corrections made afterwards. Each was applied to every affected output, and all reported numbers use the corrected outputs. The artifact we will release records model and tokenizer revisions, adapter and circuit hashes, seeds, and machine-readable results for every reported student, and every student adapter matches the hash recorded when it was trained.

Table 27: Implementation corrections made after the settings were fixed. Each error was corrected and the listed outputs were recomputed with the specified settings, so every reported result uses the settings stated in the text.

Correction	Outputs recomputed
DIPPER ran with lexical and order diversity 40 instead of the specified 60.	All DIPPER paraphrases and the students trained on them.
DIPPER sampled with top-k 50 instead of without top-k truncation.	All DIPPER paraphrases and the students trained on them.
Qwen students inherited a repetition penalty of 1.1 from the default generation settings of their checkpoint.	All answers of Qwen students, including benchmark answers.
Rewriter inputs longer than 1,536 tokens were truncated.	308 rewritten training answers of ten unwatermarked students, and these students.
Answers of the newly trained unwatermarked students were generated with the adapters of other students.	All answers and scores of these students.
The language-model classifier cut answers by characters, and the GoEmotions classifier did not read long answers window by window.	All classifier scores.
SynthID-Text, TextSeal, and Gumbel-max scoring started one position earlier than in the reference detectors.	All statistics of these three methods.
The ADFP detector used a default seed instead of the key used for generation.	All ADFP statistics.
TextSeal lacked the entropy weighting of its reference detector when the student returns its most probable next token.	TextSeal results of Table 19.
Filtered students were compared with unfiltered instead of equally filtered unwatermarked students.	Perplexity ratios of filtered students.