WMMD-BENCH: DO LLM OWNERSHIP WATERMARKS SURVIVE MODEL DISTILLATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern LLM watermarking and fingerprinting methods are increasingly used to protect model ownership, yet their robustness to model distillation remains insufficiently understood. In a model extraction setting, an adversary can distill a protected teacher into a fresh student using only teacher supervision, potentially weakening or removing the ownership evidence embedded in the original model. We introduce WMMD-Bench, a systematic evaluation of ten recent active and passive ownership mechanisms under direct hard-label distillation, semantic-preserving response transformation, online soft-logit distillation, and same- and cross-lineage student models. Across native detection metrics, clean controls, utility evaluation, applicability audits, and five 11-checkpoint training trajectories, we find that the evaluated active watermarks are not reliably preserved under model distillation. Their registered signals are substantially weakened or remain absent in distilled students, while richer soft-logit supervision provides only limited and method-dependent recovery. Training trajectories further suggest that this vulnerability is primarily associated with limited watermark acquisition, rather than strong acquisition followed by later forgetting. Passive fingerprints appear substantially more stable, but all five already produce positive detections on clean same-lineage students before distillation, indicating that their apparent robustness can be confounded by pre-existing model similarity. Overall, our results reveal model distillation as a substantial weakness of current active LLM watermarking methods and highlight the need for ownership mechanisms explicitly designed to survive knowledge transfer into independently trained student models. Code and evaluation artifacts are available at https://anonymous.4open.science/r/WMMD-Bench-FCA6/.

1 INTRODUCTION

Large language models (LLMs) have motivated growing interest in model provenance and ownership protection. Existing approaches include *active* mechanisms that deliberately embed verifiable signals or behaviors (Gloaguen et al., 2026; Nasery et al., 2025; Xiong et al., 2026; Xu et al., 2024, 2025a,b), and *passive* fingerprints that rely on intrinsic model characteristics (Hu et al., 2026; Shao et al., 2026; Zeng et al., 2024, 2026; Zhang et al., 2025). Related text-watermarking work likewise shows that detectable statistical signals can be embedded in LLM outputs (Kirchenbauer et al., 2023, 2024; Kuditipudi et al., 2024). However, whether such ownership evidence remains reliable after the protected model is used to train a new model is much less clear.

Model distillation poses a direct challenge to model ownership protection. By querying a protected teacher, an adversary can train a separately initialized student to reproduce its behavior without copying its parameters (Hinton et al., 2015; Tramèr et al., 2016). Recent LLM distillation methods further show that substantial teacher knowledge can be transferred through response- or distribution-level supervision (Agarwal et al., 2024; Gu et al., 2024b; Ko et al., 2024). This raises a central question: *can existing ownership signals survive model distillation?*

Prior studies suggest that watermark signals may behave very differently under distillation. Watermark-related biases in generated data can be learned by downstream models (Gu et al., 2024a; Sander et al., 2024), while Pan et al. (2025) show that paraphrasing and other countermeasures can substantially weaken watermark inheritance during unauthorized distillation. Recent antidistillation

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.

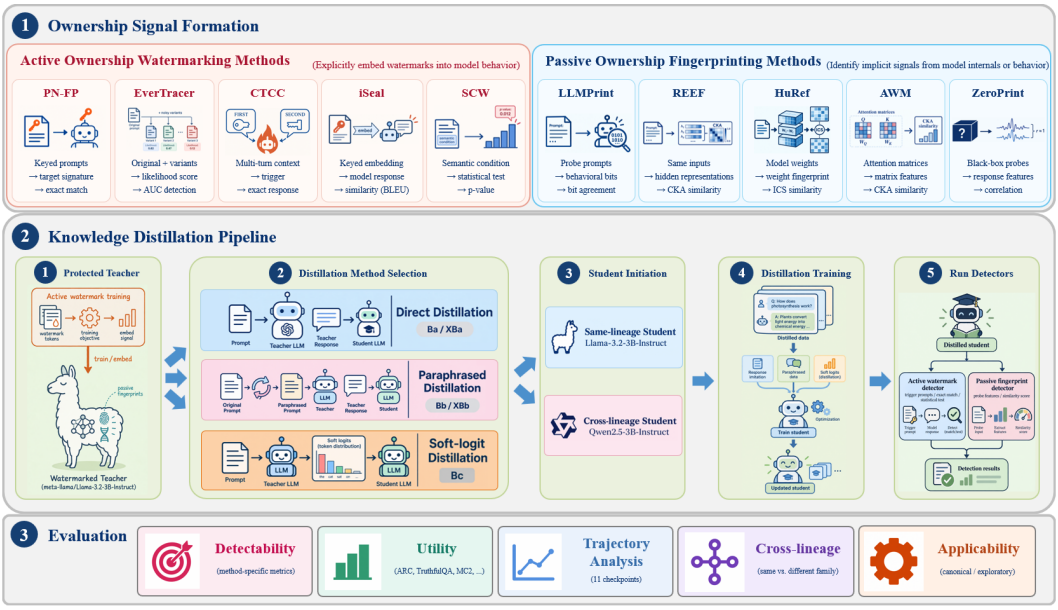

Figure 1: **Overview of WMMD-Bench.** Active watermarks and passive fingerprints are evaluated through ownership signal formation, model distillation, and ownership verification under direct (Ba/XBa), transformed-response (Bb/XBb), and soft-logit (Bc) distillation settings.

fingerprinting further explores signals explicitly designed to transfer to distilled students (Xu et al., 2026). However, existing studies largely focus on specific watermark constructions or attack settings, leaving the robustness of broader LLM ownership mechanisms under a common distillation framework insufficiently understood.

Evaluating distillation robustness requires more than endpoint detection. For active watermarks, a weak final signal may reflect either limited acquisition or subsequent forgetting, which endpoint measurements alone cannot distinguish. For passive fingerprints, positive post-distillation detection may instead arise from pre-existing similarity between the teacher and student models. We therefore include clean student controls and training trajectories to better interpret ownership evidence after distillation.

To systematically evaluate these questions, we introduce **WMMD-Bench**, covering ten recent LLM ownership mechanisms, including five active watermarks and five passive fingerprints. We evaluate hard-label, semantic-preserving transformed-response, and online soft-logit distillation with both same- and cross-lineage students. WMMD-Bench further incorporates clean student controls and eleven-checkpoint training trajectories to examine both post-distillation detectability and watermark acquisition during training.

Our contributions are threefold. ❶ We introduce **WMMD-Bench**, a systematic benchmark of ten recent LLM ownership mechanisms across three distillation settings and both same- and cross-lineage student models. ❷ We show that the evaluated active watermarks are substantially vulnerable to model distillation: their ownership signals are strongly weakened or absent in distilled students, while richer teacher supervision provides only limited and method-dependent recovery. ❸ We go beyond endpoint detection by incorporating clean controls and training trajectories, revealing that active-watermark failures are largely associated with limited signal acquisition, while the apparent robustness of passive fingerprints can be confounded by pre-existing same-lineage similarity.

2 Related Work

Existing LLM watermarking work mainly studies detectability, generation quality, and robustness to output transformations (Hou et al., 2024; Kirchenbauer et al., 2023, 2024; Kuditipudi et al., 2024; Pan et al., 2024; Piet et al., 2025; Tu et al., 2024; Wu et al., 2026; Zhao et al., 2024). Recent studies

further examine watermark stealing, learnability, radioactivity, and inheritance under unauthorized distillation (Gu et al., 2024a; Jovanovic et al., 2024; Pan et al., 2025; Sander et al., 2024), while ADFP explicitly designs signals for acquisition by distilled students (Xu et al., 2026). WMMD-Bench broadens this evaluation to five active mechanisms (PN-FP, EverTracer, CTCC, iSeal, and SCW) and five passive fingerprints (LLMPrint, REEF, HuRef, AWM, and ZeroPrint), preserving each method's native verification metric. We evaluate these mechanisms across supervision type, clean initialization, training dynamics, student lineage, and detector applicability; detailed method descriptions and prior-work comparisons are provided in Appendix J.1.

3 WMMD-Bench Framework

WMMD-Bench is designed to systematically evaluate whether LLM ownership evidence survives unauthorized model distillation. The benchmark considers three main dimensions: ownership mechanism, distillation supervision, and student lineage. We further use complementary evaluation evidence to distinguish post-distillation detectability from signal acquisition and pre-existing model similarity. Figure 1 illustrates the overall framework.

3.1 Problem Setup and Threat Model

Let T denote a protected teacher model and S_{θ_0} a fresh student initialized independently of the protected teacher. Given an extraction query set $\mathcal{X} = \{x_i\}_{i=1}^{N}$, an adversary obtains supervision from T through a distillation channel $\mathcal{Q}_c$, where c specifies the available teacher information. The distilled student is

$$S_{\theta_c^*} = \mathcal{A}(S_{\theta_0}, \mathcal{Q}_c(T, \mathcal{X}))\,, \tag{1}$$

where $\mathcal{A}$ denotes the student-training procedure.

We consider this process as an unauthorized model-extraction attack against ownership protection: the adversary trains a new student from teacher supervision without directly copying or modifying the teacher parameters. Depending on the attack setting, $\mathcal{Q}_c$ exposes teacher responses, transformed responses, or predictive distributions.

For an ownership mechanism m, let

$$s_m(M; T) = \mathcal{V}_m(M, T) \tag{2}$$

denote the native ownership score produced by verifier $\mathcal{V}_m$ for a candidate model M with respect to the protected teacher or its registered ownership evidence.

For active watermarks, robustness requires the deliberately embedded ownership signal to remain detectable in $S_{\theta_c^*}$. For passive fingerprints, a positive post-distillation score may additionally arise from model similarity already present in S_{θ_0}. WMMD-Bench therefore evaluates both clean and distilled students.

3.2 Benchmark Composition

WMMD-Bench organizes each evaluation condition along three dimensions:

$$\mathcal{B} = \mathcal{M} \times \mathcal{C} \times \mathcal{L}, \tag{3}$$

where $\mathcal{M}$ denotes the ownership mechanism, $\mathcal{C}$ the distillation supervision, and $\mathcal{L}$ the student lineage.

Ownership mechanisms. We evaluate ten recent ownership mechanisms, consisting of five active watermarks and five passive fingerprints. Active methods deliberately introduce registered ownership signals into the protected teacher, whereas passive methods identify models from intrinsic behavioral, representational, or parameter-level characteristics. Each method retains its original verification interface and native detector metric.

Distillation supervision. We consider three forms of teacher supervision. Direct response distillation (Ba) trains the student on ordinary teacher responses. Transformed-response distillation (Bb) applies a frozen semantic-preserving transformation before student training, testing robustness to altered surface realization. Soft-logit distillation (Bc) provides richer distribution-level teacher supervision and tests whether increased teacher information strengthens ownership-signal transfer.

Student lineage. We evaluate both same-lineage and cross-lineage students. Same-lineage students share the teacher model family, whereas cross-lineage students use a different model family. Cross-lineage variants of Ba and Bb are denoted XBa and XBb. Because Bc requires compatible token identities and vocabularies, it is evaluated only in the same-lineage setting.

Cross-lineage comparison is used as a robustness probe rather than a causal ablation of lineage, since changing model family also changes architecture, tokenization, and pretraining.

3.3 Evaluation

A single post-distillation detector score is insufficient to characterize ownership robustness. WMMD-Bench therefore combines several complementary forms of evidence.

Endpoint detection. We apply each method's native verifier to the distilled student to determine whether ownership evidence remains detectable after extraction.

Clean control. The same verifier is applied to the student before distillation. This provides a baseline for identifying ownership evidence that is already present without teacher supervision.

Training dynamics. For active watermarks under direct same-lineage distillation, we evaluate eleven checkpoints throughout training. These trajectories distinguish weak signal acquisition from successful acquisition followed by subsequent forgetting.

Utility and applicability. We additionally measure student utility and verify whether each detector remains meaningful under the evaluated model family. Utility helps rule out general model collapse, while applicability prevents invalid cross-model detector outputs from being interpreted as watermark failure.

Together, these measurements evaluate not only whether ownership evidence remains detectable after distillation, but also how that evidence emerges and whether it can be meaningfully interpreted.

4 Experimental Setup

4.1 Models, Data, and Training

The main benchmark uses Llama-3.2-3B-Instruct as the same-lineage student and Qwen2.5-3B-Instruct as the cross-lineage student (Grattafiori et al., 2024; Qwen et al., 2024). Each student starts independently from an unmodified instruction-tuned checkpoint; no distillation condition continues training from another condition.

For each method-specific teacher, we use a fixed extraction budget of 20,000 QA records. Unless otherwise specified, students are trained with full-parameter BF16 optimization for three epochs using an effective batch size of eight, a learning rate of 10^{-5}, and seed 42, resulting in 7,500 optimizer steps.

Across direct-response, transformed-response, and soft-logit distillation, we match the record budget, optimization horizon, effective batch size, learning rate, seed, and response-only masking whenever objective compatibility permits. Equal record budgets do not necessarily imply identical data or token exposure across methods; the PN-FP and SCW follow-up experiments use protocol-matched reconstructed parents.

Appendix B provides exact model revisions, dataset identities, optimization details, and serialization rules.

4.2 Distillation Implementation

Direct response distillation (Ba/XBa). Students are trained on ordinary responses generated by the protected teacher using the fixed 20,000-record extraction budget. Training uses response-only supervision, with prompt positions excluded from the loss. Ba uses the same-lineage Llama student, while XBa uses the cross-lineage Qwen student.

Transformed-response distillation (Bb/XBb). Teacher responses are rewritten by Qwen2.5-3B-Instruct before student training while preserving the instruction, sample identity, and intended meaning. The transformation is frozen before distillation and is detector-agnostic, with no watermark-specific optimization (Pan et al., 2025). Atomic answers and unsuccessful rewrites are retained unchanged. Appendix C provides the prompt, sampling procedure, filtering rules, and dataset statistics.

Soft-logit distillation (Bc). Bc provides distribution-level teacher supervision in addition to response targets. At assistant-response positions $\mathcal{M}$, we optimize

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left[0.5\, \mathrm{CE}(y_t, p_S^{(1)}) + 0.5\, \tau^2 D_{\mathrm{KL}}\left(p_T^{(\tau)} \| p_S^{(\tau)} \right) \right], \qquad \tau = 2, \tag{4}$$

where $p_T^{(\tau)}$ and $p_S^{(\tau)}$ denote the temperature-scaled teacher and student distributions. The teacher remains frozen and prompt positions are masked. Bc uses 225 warmup steps followed by cosine learning-rate decay.

We evaluate Bc for all five active mechanisms and the shared passive-student condition. Because logit distillation requires aligned token identities and vocabulary dimensions, Bc is evaluated only with the same-lineage Llama student.

4.3 Ownership and Utility Evaluation

Ownership verification. Each method is evaluated using its original verification interface, native score, and frozen decision rule. We report measurements for the protected teacher/reference, clean students, and distilled students. For the five active methods under direct same-lineage distillation, we additionally evaluate eleven training checkpoints. Method-specific detector definitions, thresholds, and implementation deviations are provided in Appendix A.

Cross-lineage applicability. Changing the student model family does not always preserve the semantics of the original detector. PN-FP, CTCC, and SCW retain operational Qwen-compatible verification interfaces. EverTracer is reported as N/A under canonical cross-lineage evaluation because tokenization changes the underlying token-level measurement, while iSeal is N/A because applying its keyed interface to Qwen would require redefining the original verifier. Exploratory measurements, where reported, are kept separate from canonical benchmark results.

Utility evaluation. We evaluate student utility using ARC-Challenge normalized accuracy and TruthfulQA MC2 (Clark et al., 2018; Lin et al., 2022). Cross-lineage utility changes are computed relative to the clean Qwen student, while same-lineage values are reported in their corresponding clean-Llama evaluation contexts. Full results and evaluation settings are provided in Appendix D.

All experiments were conducted on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU using BF16 precision; complete software and hardware details are provided in Appendix B.

5 Results

Table 1 summarizes WMMD-Bench results. We first examine overall ownership robustness, then analyze acquisition dynamics, supervision type, model lineage, and additional validity checks.

The eight reported conditions do not correspond one-to-one to independent training runs: reference and non-applicable cells are included, and the five passive methods share the same student within each common distillation condition.

Table 1: **Native ownership-verification results across WMMD-Bench.** PN-FP, CTCC, and iSeal report counts out of 1,024, 95, and 200 probes; EverTracer reports member-oriented AUC; SCW reports a p-value, where smaller is stronger. Remaining methods retain their native similarity metrics, which are not directly comparable across methods. N/A denotes non-applicable and EXP an exploratory diagnostic. Full-precision values and decision rules are provided in Appendix A.

	Reference	Same-lineage Llama				Cross-lineage Qwen		
Method	Teacher	Clean L	Ba	Bb*	Bc	Clean Q	XBa	XBb*
PN-FP	956	1	98	72	110	43	84	67
EverTracer	1.000	0.442	0.499	0.476	0.590	N/A	N/A (EXP)	N/A (EXP)
CTCC	95	0	0	0	0	0	0	0
iSeal	179	0	0	0	0	N/A	N/A	N/A
SCW	0.000	0.920	0.844	0.953	0.882	0.745	0.916	0.984
LLMPrint	1.000	1.000	0.820	0.785	0.860	0.680	0.680	0.695
REEF	1.000	1.000	0.954	0.967	0.986	0.455	0.471	0.473
HuRef	100.0	100.0	100.0	100.0	100.0	-8.4	-8.4	-8.3
AWM	1.000	1.000	1.000	1.000	1.000	0.002	0.002	0.002
ZeroPrint	1.000	1.000	0.840	0.917	0.927	0.679	0.702	0.702

5.1 Overall ownership robustness under distillation

Across the five active mechanisms, Table 1 shows a substantial gap between protected teachers and distilled students. PN-FP produces 956/1,024 matching signatures in the teacher, compared with 1/1,024 in clean Llama and 98/1,024 after direct-response distillation (Ba). EverTracer drops from teacher AUC 1.0 to near chance after hard-label distillation. CTCC produces no triggered activations among 95 probes, iSeal no registered successes among 200 keyed queries for same-lineage students, and SCW remains above its $\alpha = 0.001$ detection threshold.

Thus, the evaluated active ownership signals are not reliably preserved under model distillation. Endpoint measurements alone, however, cannot distinguish weak acquisition from successful acquisition followed by forgetting; we examine this distinction in Section 5.2.

Passive fingerprints show a different pattern. All five are already positive on the clean same-lineage student and remain positive after Ba, Bb, and Bc. Post-distillation detectability therefore does not by itself establish newly transferred fingerprint information. We examine this lineage effect in Section 5.4.

5.2 Active watermark trajectories indicate limited acquisition

Endpoint non-detection cannot distinguish a watermark that was never substantially acquired from one acquired and later forgotten. The eleven-checkpoint trajectories in Figure 2 provide additional evidence.

PN-FP rises from 1 match at initialization to 3 at step 50, 58 at step 100, 79 at step 250, and 99 by step 4,000, remaining there through the final checkpoint. This is only about one tenth of the teacher's 956 signatures. Most acquisition occurs early, followed by a low plateau; the small decrease from 86 to 85 does not indicate substantial collapse. The trajectory is therefore consistent with partial acquisition rather than strong acquisition followed by forgetting.

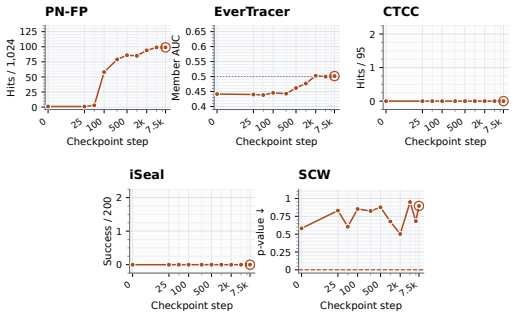

Figure 2: Active ownership trajectories under direct same-lineage distillation. Eleven checkpoints are measured per method; step zero is clean Llama. EverTracer chance AUC is 0.5, and the SCW threshold is $\alpha = 0.001$. The x-axis is symlog and linear through step 25. Full results are in Appendix E.

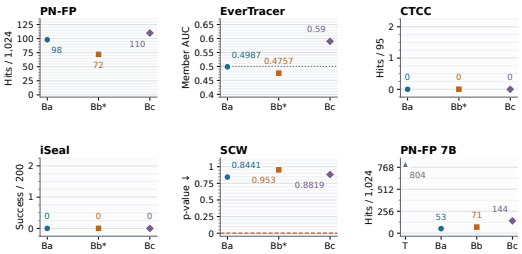

Figure 3: Active ownership signals across distillation supervision. Native metrics are shown for the five 3B active methods and the representative PN-FP 7B extension. Ba, Bb*, and Bc denote direct-response, transformed-response, and soft-logit distillation. Comparisons are descriptive single-seed endpoints.

The other active methods show no general hidden acquisition-and-collapse pattern. EverTracer remains near chance, CTCC and iSeal never produce nonzero registered successes, and SCW remains non-significant under $\alpha = 0.001$. Across all five methods, the measured trajectories are more consistent with limited acquisition than later loss. This conclusion is restricted to sampled checkpoints under same-lineage Ba; transient behavior and dynamics under Bb, Bc, XBa, and XBb are not observed.

5.3 Changing distillation supervision does not reliably restore active watermarks

Figure 3 compares active ownership signals across the three same-lineage supervision settings. Changing supervision alters some native measurements but does not consistently restore active ownership evidence.

Soft-logit supervision. Bc yields limited and method-dependent recovery. PN-FP reaches 110/1,024 matches versus 98/1,024 under Ba, while EverTracer increases from AUC 0.4987 to 0.59. These are descriptive rather than significance-tested improvements. CTCC and iSeal remain at zero, and SCW remains undetected. Richer teacher information can therefore strengthen some measurements without reliably preserving active watermarks.

Transformed-response supervision. The effect of semantic-preserving transformation is also method-dependent. PN-FP decreases from 98 under Ba to 72 under Bb, and from 84 under XBa to 67 under XBb. Other methods show no consistent attenuation pattern. Because atomic answers and unsuccessful rewrites may remain unchanged, Bb denotes a transformed-response condition rather than a guarantee that every response was paraphrased.

Overall, neither soft-logit supervision nor transformed responses change the main conclusion: supervision affects some detector measurements but does not reliably restore active ownership signals.

5.4 Passive fingerprint detectability is strongly associated with model lineage

Figure 4 shows that all five passive fingerprints are already positive on clean same-lineage Llama. LLMPrint scores 1.0 before distillation and 0.82, 0.785, and 0.86 after Ba, Bb, and Bc; ZeroPrint similarly changes from 1.0 to 0.8399, 0.9170, and 0.9270. REEF, HuRef, and AWM also remain near their reference-score regions. Positive post-distillation detection therefore does not establish newly acquired fingerprint information.

Cross-lineage Qwen students occupy substantially different native-score regions, indicating a strong association between passive detectability and model relatedness. However, replacing Llama with Qwen jointly changes model family, architecture, tokenization, pretraining, and representation geometry, so lineage is not causally isolated.

REEF, AWM, and ZeroPrint nevertheless exceed their Qwen-calibrated thresholds. Because Qwen participates in threshold calibration, these crossings are not independent held-out evidence of distillation-induced transfer. Clean initialization is therefore essential when interpreting apparent passive fingerprint retention.

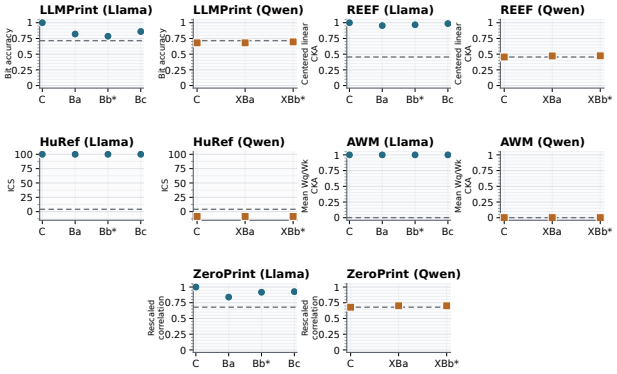

Figure 4: **Passive fingerprint scores across same- and cross-lineage students.** Each method retains its native metric; dashed lines indicate detector thresholds. C denotes the clean student, and XBa/XBb use Qwen. All five passive fingerprints are already positive on clean same-lineage Llama.

5.5 Additional validity and robustness checks

Signal-exposure audit. The ordinary distillation datasets contain no exact registered or full-string matches under the method-specific checks for PN-FP, EverTracer, CTCC, iSeal, or SCW. Because these checks operate at different granularities, zero exact overlap does not exclude shared tokens, semantic exposure, indirect representation learning, teacher-distribution information, or key-dependent features. The audit is therefore consistent with limited direct exposure but does not explain why active signals fail to transfer.

Table 2: **Representative 7B robustness extension.** PN-FP and AWM are evaluated under Ba, Bb, and Bc using native metrics. This is a robustness check rather than a controlled scale ablation.

Method	Teacher / Ref.	Ba	Bb	Bc
PN-FP	804/1,024	53/1,024	71/1,024	144/1,024
AWM	1.000000	0.999995	0.999993	0.999996

Cross-lineage detector applicability. EverTracer yields exploratory Qwen AUCs of 0.523 for XBa and 0.5227 for XBb, but its canonical cross-lineage cells remain N/A because tokenization changes the measured quantity. Applying iSeal to Qwen would likewise require redefining its keyed interface. These are verifier-portability limitations rather than student-training failures; Appendix G provides details.

7B extension. For PN-FP, the 7B teacher produces 804/1,024 matches, compared with 53/1,024 after Ba, 71/1,024 after Bb, and 144/1,024 after Bc; the largest student count is only 17.9% of the teacher value. AWM remains near its reference similarity across all three conditions. Because only PN-FP and AWM are evaluated at 7B, this is a robustness extension rather than a controlled scaling result.

Student utility. The distilled students remain evaluable on both utility benchmarks, weakening a whole-model-collapse explanation for poor active-watermark retention. Qwen shows task-dependent changes; for example, the shared passive XBa student improves ARC-Challenge while reducing TruthfulQA MC2 relative to clean Qwen (Appendix Figure A1). Two tasks, however, do not establish general capability preservation or exclude narrower behavioral degradation.

6 Discussion

6.1 Interpreting ownership robustness

Active and passive ownership mechanisms exhibit fundamentally different behaviors under distillation. For active watermarks, the sampled trajectories are more consistent with limited acquisition

than with strong acquisition followed by forgetting. Ordinary distillation reproduces teacher behavior but does not necessarily expose or optimize the watermark's registered verification interface. The exposure audit is consistent with this interpretation, although it does not establish a causal link between signal exposure and acquisition.

Passive fingerprints require a different interpretation. All five are already positive on the clean same-lineage student, so high post-distillation scores cannot by themselves be attributed to information transferred from the protected teacher. Cross-lineage results further associate passive detectability with model relatedness, although model family, architecture, tokenization, and pretraining change jointly and are not causally isolated. Likewise, richer soft-logit supervision improves some measurements but does not reliably restore active signals, suggesting that robustness depends not only on how much teacher information is transferred, but also on whether that information aligns with the ownership mechanism.

6.2 Implications for ownership-mechanism design

Ownership mechanisms intended to survive model extraction should jointly consider *acquisition*, *specificity*, and *verifier portability*. The ownership signal should be expressed in behavior a realistic extractor is likely to learn, remain specific to the protected model rather than shared lineage or initialization, and retain a meaningful verification interface across plausible student architectures and tokenizers.

This creates a tension between transferability and specificity. Rare or highly structured signals may be easy to verify in the teacher but difficult to acquire through distillation, whereas signals tied to general model geometry may remain detectable across related models while providing weaker evidence of transfer from a particular teacher. A promising direction is therefore to design signals that occur naturally in ordinary semantic or behavioral outputs while maintaining low false-positive rates.

6.3 Limitations and future directions

Most experiments use a single seed and 3B models; the 7B extension covers only PN-FP and AWM and is not a controlled scale ablation. Trajectories are available only for same-lineage active methods under direct distillation, while Bc is endpoint-only. Cross-lineage comparisons change several model properties simultaneously, Qwen participates in calibration for several passive detectors, and utility is evaluated on only two downstream tasks.

Future work should include repeated seeds, additional model families and scales, and trajectories under transformed-response, soft-logit, and cross-lineage settings. Controlled manipulation of registered-signal exposure could directly test whether limited acquisition results from insufficient exposure, while separately varying tokenizer, architecture, and initialization would help disentangle lineage effects from detector portability.

7 Conclusion

WMMD-Bench provides a systematic evaluation of ten LLM ownership mechanisms under model distillation. Across the evaluated settings, active watermarks are not reliably preserved: their registered signals are substantially weakened or remain absent in distilled students, while richer soft-logit supervision provides only limited and method-dependent recovery. Training trajectories further suggest that this failure is more consistent with limited signal acquisition than with strong acquisition followed by subsequent forgetting. Passive fingerprints show a different behavior: they often remain detectable after distillation, but all five are already positive on clean same-lineage students, making post-distillation detection insufficient by itself to establish newly transferred ownership information.

These findings show that robustness to distillation requires more than strong verification on the protected teacher. Reliable ownership mechanisms must support signal acquisition through realistic extraction channels while remaining specific to the protected model and verifiable across plausible student architectures. Designing ownership evidence that jointly achieves acquisition, specificity, and verifier portability remains an important challenge for practical LLM ownership protection.

REPRODUCIBILITY STATEMENT

The appendix documents model revisions, training and detector configurations, dataset identities, utility settings, and checkpoint observations. The accompanying anonymous supplementary project includes the plotted and tabulated result data, figure/table-generation code, and content hashes for recorded assets. These materials support inspection and reproduction of the reported analyses. Model weights, full dataset contents, and complete raw experiment logs are not included in the submission package.

ETHICS STATEMENT

Ownership verification can support legitimate provenance claims, but model extraction and fingerprint removal are dual-use topics. We study bounded, reproducible transformations of research models and do not assert legal ownership from detector outputs. False positives and invalid cross-model comparisons can harm both model developers and independent users. We therefore preserve clean controls, distinguish calibration from held-out evaluation, and report non-applicable detectors explicitly. We do not release private credentials or operational infrastructure in this submission package.

AI USE STATEMENT

AI assistance was used for organizing the literature, drafting and revising prose, preparing LaTeX and visualization code, and checking consistency against existing experiment artifacts. It was not used to invent experimental measurements. The authors remain responsible for the accuracy, originality, citations, interpretation, and final verification of the submitted manuscript and supplementary material.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *ICLR*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

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. URL https://arxiv.org/abs/1803.05457.

Thibaud Gloaguen, Robin Staab, Nikola Jovanovic, and Martin Vechev. LLM Fingerprinting via Semantically Conditioned Watermarks. In *ICLR*, 2026. URL https://arxiv.org/abs/2505.16723.

Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chenchen Gu, Xiang Lisa Li, Percy Liang, and Tatsunori Hashimoto. On the Learnability of Watermarks for Language Models. In *ICLR*, 2024a. URL https://arxiv.org/abs/2312.04469.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *ICLR*, 2024b. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

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 *NAACL*, 2024. URL https://aclanthology.org/2024.naacl-long.226/.

Yuepeng Hu, Zhengyuan Jiang, Mengyuan Li, Osama Ahmed, Zhicong Huang, Cheng Hong, and Neil Zhenqiang Gong. Fingerprinting LLMs via Prompt Injection. In *ACL*, 2026. URL https://aclanthology.org/2026.acl-long.541/.

Nikola Jovanovic, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/jovanovic24a.html.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. In *ICML*, 2023. URL https://arxiv.org/abs/2301.10226.

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 *ICLR*, 2024. URL https://arxiv.org/abs/2306.04634.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards Streamlined Distillation for Large Language Models. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust Distortion-free Watermarks for Language Models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL*, 2022. URL https://arxiv.org/abs/2109.07958.

Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. Scalable Fingerprinting of Large Language Models. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2502.07760.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An Open-Source Toolkit for LLM Watermarking. In *EMNLP: System Demonstrations*, 2024. URL https://aclanthology.org/2024.emnlp-demo.7/.

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 *ACL*, 2025. URL https://arxiv.org/abs/2502.11598.

Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. MARKMyWORDS: Analyzing and Evaluating Language Model Watermarks. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025. URL https://ieeexplore.ieee.org/document/10992530/.

Qwen et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking Makes Language Models Radioactive. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2402.14904.

Shuo Shao, Yiming Li, Hongwei Yao, Yifei Chen, Yuchen Yang, and Zhan Qin. Reading Between the Lines: Towards Reliable Black-box LLM Fingerprinting via Zeroth-order Gradient Estimation. In *The ACM Web Conference*, 2026. URL https://arxiv.org/abs/2510.06605.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *25th USENIX Security Symposium*, pp. 601–618, 2016. URL https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.83/.

Yihan Wu, Xuehao Cui, Ruibo Chen, and Heng Huang. Analyzing and Evaluating Unbiased Language Model Watermark. In *ICLR*, 2026. URL https://arxiv.org/abs/2509.24048.

Zixun Xiong, Gaoyi Wu, Qingyang Yu, Mingyu Derek Ma, Lingfeng Yao, Miao Pan, Xiaojiang Du, and Hao Wang. iSeal: Encrypted Fingerprinting for Reliable LLM Ownership Verification. In *AAAI*, 2026. URL https://arxiv.org/abs/2511.08905.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional Fingerprinting of Large Language Models. In *NAACL*, 2024. URL https://aclanthology.org/2024.naacl-long.180/.

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. URL https://arxiv.org/abs/2602.03812.

Zhenhua Xu, Meng Han, and Wenpeng Xing. EverTracer: Hunting Stolen Large Language Models via Stealthy and Robust Probabilistic Fingerprint. In *EMNLP*, 2025a. URL https://arxiv.org/abs/2509.03058.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. CTCC: A Robust and Stealthy Fingerprinting Framework for Large Language Models via Cross-Turn Contextual Correlation Backdoor. In *EMNLP*, 2025b. URL https://arxiv.org/abs/2509.09703.

Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. HuRef: HUman-REadable Fingerprint for Large Language Models. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2312.04828.

Boyi Zeng, Lin Chen, Ziwei He, Xinbing Wang, and Zhouhan Lin. AWM: Accurate Weight-Matrix Fingerprint for Large Language Models. In *ICLR*, 2026. URL https://arxiv.org/abs/2510.06738.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. REEF: Representation Encoding Fingerprints for Large Language Models. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.14273.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable Robust Watermarking for AI-Generated Text. In *ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/beae9ed5316bcc48e616754c06c11875-Abstract-Conference.html.

APPENDIX OVERVIEW

The appendix provides the implementation details and supporting analyses underlying the main results. Appendix A defines the ownership mechanisms, native verification metrics, and detector applicability used throughout the benchmark. Appendices B and C describe the experimental protocol and the three distillation conditions in detail, including model initialization, data construction, optimization, response transformation, and soft-logit supervision. Appendix D reports the complete endpoint and utility results, while Appendices E–G expand the analyses of watermark acquisition, signal exposure, and cross-lineage detector validity. Appendix H synthesizes these results across ownership mechanism, supervision type, and student lineage and develops the corresponding design implications. Appendix I presents the representative 7B extension, followed by comparison with related evaluation settings and reproducibility notes in the final appendices.

A OWNERSHIP MECHANISMS AND VERIFICATION

The ten ownership mechanisms comprise five active watermarks and five passive fingerprints. Their verification procedures are heterogeneous: some operate on registered query–response behavior, while others compare likelihoods, representations, model parameters, or response-derived features. WMMD-Bench therefore retains each method's native verification interface and metric rather than converting them to a common ownership score.

For a candidate model M, we write the native score of ownership mechanism m as

$$s_m(M) = V_m(M; \mathcal{R}_m),$$

where V_m is the method-specific verifier and $\mathcal{R}_m$ denotes the registered probes, reference model information, or other fixed evidence required by that verifier. When the original method defines a decision rule, we write

$$d_m(M) = D_m(s_m(M); \mathcal{R}_m)\,.$$

The form of s_m and D_m remains method-specific: a score may be a match count, an AUC, a p-value, or a similarity measure, and some active methods define success per probe without an additional model-level cutoff. Table A1 summarizes these interfaces. Complete endpoint values corresponding to Table 1 are reported in Appendix D.1.

A.1 PN-FP

PN-FP embeds registered behavioral signatures generated through Perinucleus sampling (Nasery et al., 2025). Let

$$\mathcal{R}_{\mathrm{PN}} = \{(k_i, t_i)\}_{i=1}^{N_{\mathrm{PN}}}, \qquad N_{\mathrm{PN}} = 1024,$$

denote the fixed secret key–target pairs, and let $r_M(k_i)$ be the response generated by candidate model M for key k_i. WMMD-Bench measures

$$s_{\mathrm{PN}}(M) = \sum_{i=1}^{N_{\mathrm{PN}}} \mathbf{1}[r_M(k_i) = t_i]\,, \qquad \mathrm{rate}_{\mathrm{PN}}(M) = \frac{s_{\mathrm{PN}}(M)}{N_{\mathrm{PN}}},$$

where equality follows the registered complete-signature rule.

The key set, serialization, and matching rule are fixed across evaluations. We do not introduce an additional model-level ownership threshold; individual registered matches are reported directly. A nonzero clean-control count is therefore retained as a measured baseline rather than interpreted by itself as ownership evidence. Cross-lineage evaluation uses the same registered text keys and targets without re-registering them for Qwen.

A.2 EVERTRACER

EverTracer embeds probabilistic fingerprints through memorization of natural texts and verifies them through likelihood-based membership evidence (Xu et al., 2025a). For a tokenized text $x = (x_1, \ldots, x_L)$, define its mean next-token loss under candidate model M as

$$\ell_M(x) = -\frac{1}{L-1} \sum_{t=2}^{L} \log p_M(x_t \mid x_{<t}),$$

Table A1: Ownership mechanisms and verification interfaces used in WMMD-Bench. "Per probe" denotes a query-level success criterion without an additional model-level cutoff. EverTracer retains its empirical ROC operating procedure. Method-specific thresholds and cross-lineage interpretation are described below.

Category	Method	Interface / access	Native metric	Decision rule	Cross-lineage
Active watermark	PN-FP	Secret text probes; responses	Matches / 1,024	Complete signature; per probe	Canonical
	EverTracer	Fixed texts; token likelihoods	Member-oriented AUC	ROC; FPR ≤ 0.05	Exploratory only
	CTCC	Contextual probes; responses	Trigger count / 95	Exact target; per probe	Canonical
	iSeal	Keyed embedding access	Successes / 200	BLEU ≥ 50; per probe	N/A
	SCW	Semantic queries; responses	p-value	$p < 0.001$	Canonical
Passive fingerprint	LLMPrint	Behavioral probes; responses	Bit accuracy	Score $\geq$ threshold	Canonical
	REEF	Hidden representations	Centered linear CKA	Score $>$ threshold	Canonical
	HuRef	Model parameters	ICS	Score $>$ threshold	Canonical
	AWM	Attention query/key weights	Unbiased linear CKA	Score $>$ threshold	Canonical
	ZeroPrint	External probes; responses	Rescaled correlation	Score $>$ threshold	Canonical

and the corresponding likelihood score as

$$q_M(x) = \exp[-\ell_M(x)] .$$

Given the fixed registered member and nonmember sets, $\mathcal{D}_{\text{mem}}$ and $\mathcal{D}_{\text{non}}$, the primary WMMD-Bench endpoint is

$$s_{\text{ET}}(M) = \text{AUC}\left(\{q_M(x) : x \in \mathcal{D}_{\text{mem}}\}, \{q_M(x) : x \in \mathcal{D}_{\text{non}}\}\right),$$

with the registered member-oriented score direction.

We additionally retain the empirical TPR, FPR, and operating threshold at FPR ≤ 0.05 where available. This operating point belongs to the ROC procedure and is not treated as a universal binary ownership cutoff. Because the token sequence $(x_1, \ldots, x_L)$ and its prediction positions depend on the tokenizer, the canonical cross-lineage measurement is N/A; the adapted Qwen measurements are reported separately as exploratory results in Appendix G.1.

A.3 CTCC

CTCC constructs fingerprint behavior through contextual correlations distributed across multiple dialogue turns rather than a single token-level trigger (Xu et al., 2025b). WMMD-Bench evaluates an operational generation-based form of this verification task. Let

$$\mathcal{Q}_{\text{CTCC}} = \{q_i\}_{i=1}^{N_{\text{CTCC}}}, \qquad N_{\text{CTCC}} = 95,$$

be the fixed trigger-query set. The measured score is

$$s_{\text{CTCC}}(M) = \sum_{i=1}^{N_{\text{CTCC}}} \mathbf{1}[\text{strip}(r_M(q_i)) = \texttt{IAMALIVE}] .$$

The equality is case-sensitive; substrings and semantic variants do not count. The accompanying 100 suppression queries and 105 normal queries are evaluated as separate negative controls and are not included in s_{CTCC}. This operational measurement is used consistently throughout WMMD-Bench but is not claimed to reproduce every component of the original CTCC verifier. Cross-lineage evaluation preserves the same probes and exact-response rule for Qwen.

A.4 iSeal

iSeal combines an injected fingerprint with a registered keyed verification interface designed to remain robust under verification-time manipulation (Xiong et al., 2026). For registered query–target pair (q_i, t_i), let

$$b_i(M) = \mathrm{BLEU}(r_M(q_i), t_i)$$

denote the sentence-level BLEU score of the candidate response. With $N_{\mathrm{iSeal}} = 200$ registered queries, WMMD-Bench reports

$$s_{\mathrm{iSeal}}(M) = \sum_{i=1}^{N_{\mathrm{iSeal}}} \mathbf{1}[b_i(M) \geq 50] \,.$$

Thus BLEU ≥ 50 defines a per-query registered success rather than a new model-level threshold. An additional 100 held-out queries are evaluated separately, while mean and median BLEU provide auxiliary response-overlap statistics. The keyed embedding interface is architecture-dependent; applying it to the evaluated Qwen model would require redefining the verifier rather than executing the same registered measurement. Canonical cross-lineage iSeal verification is therefore N/A.

A.5 SCW

SCW embeds ownership evidence through semantically conditioned watermark behavior and applies a registered statistical test to generated responses (Gloaguen et al., 2026). For the fixed set of 1,000 verification queries, let p_M denote the p-value returned by the native detector. WMMD-Bench retains

$$s_{\mathrm{SCW}}(M) = p_M,$$

with decision

$$d_{\mathrm{SCW}}(M) = \mathbf{1}[p_M < \alpha] \,, \qquad \alpha = 0.001.$$

Smaller p-values provide stronger evidence under the registered test. The query set, statistical procedure, and decision threshold are fixed before student evaluation and are retained for Qwen. For models that remain above the decision threshold, differences among non-significant p-values are not interpreted as a calibrated measure of watermark strength.

A.6 Structural Characteristics of Active Watermark Signals

The five active watermarks differ not only in their native metrics, but also in the structure of the behavior or evidence required for successful verification. Table A2 summarizes these differences descriptively. The table does not define a numerical notion of watermark complexity; instead, it records the context and verification structure associated with each registered signal.

Table A2: Structural characteristics of the five active watermark signals. "Context structure" describes the information pattern required by the registered verification procedure and is not a quantitative complexity score.

Method	Registered signal	Context structure	Verification unit
PN-FP	Registered key–target behavioral association	Single-query	Exact target response for each secret key
EverTracer	Memorization signal over registered natural texts	Document-level	Likelihood-based member/nonmember separation
CTCC	Triggered behavior defined through contextual correlation	Multi-turn	Exact target behavior under the registered dialogue context
iSeal	Registered encrypted/keyed fingerprint behavior	Keyed-interface dependent	Per-query response success under the registered verifier
SCW	Semantically conditioned statistical watermark behavior	Semantic-condition dependent	Aggregate statistical evidence over the query set

These mechanisms therefore impose different verification structures on the candidate model. PN-FP is evaluated through individual registered query–response associations, whereas CTCC requires the relevant contextual relation across dialogue turns before its registered target behavior can be observed. EverTracer instead aggregates likelihood-based evidence over registered natural texts, while iSeal and SCW rely on keyed and statistical verification structures, respectively. These distinctions provide context for the acquisition trajectories and exposure measurements reported later in the appendix.

A.7 LLMPRINT

LLMPrint constructs behavioral fingerprints by optimizing prompt-injection probes that induce model-specific token preferences (Hu et al., 2026). Let

$$\mathcal{R}_{\mathrm{LP}} = \{(q_i, b_i)\}_{i=1}^{N_{\mathrm{LP}}}$$

denote the fixed probe-to-bit registration, and let $\hat{b}_i(M)$ be the bit recovered from candidate model M on probe q_i. The native bit-accuracy score is

$$s_{\mathrm{LP}}(M) = \frac{1}{N_{\mathrm{LP}}} \sum_{i=1}^{N_{\mathrm{LP}}} \mathbf{1}\left[\hat{b}_i(M) = b_i\right].$$

The frozen detector decision is

$$d_{\mathrm{LP}}(M) = \mathbf{1}[s_{\mathrm{LP}}(M) \geq \tau_{\mathrm{LP}}], \qquad \tau_{\mathrm{LP}} = 0.7150049776126003.$$

The same registered behavioral interface is evaluated on clean and distilled Llama and Qwen students. Because the clean same-lineage Llama student is already positive before distillation, positive post-distillation agreement cannot by itself establish newly transferred fingerprint information.

A.8 REEF

REEF identifies model relatedness by comparing internal representations of a reference and candidate model on the same inputs (Zhang et al., 2025). Let H_T and H_M denote the fixed extracted representation matrices of the registered reference T and candidate model M, respectively. WMMD-Bench retains the centered linear CKA score

$$s_{\mathrm{REEF}}(M) = \mathrm{CKA}_{\mathrm{lin}}(H_T, H_M),$$

using the fixed feature extraction, layer selection, centering, and normalization procedure. The detector decision is

$$d_{\mathrm{REEF}}(M) = \mathbf{1}[s_{\mathrm{REEF}}(M) > \tau_{\mathrm{REEF}}], \qquad \tau_{\mathrm{REEF}} = 0.4546738923165847.$$

The clean-Qwen score equals this strict threshold because Qwen participates in the detector calibration. Cross-lineage scores remain valid outputs of the specified representation comparison, but threshold crossings involving this calibration model are not independent held-out evidence of fingerprint transfer. Appendix G.3 discusses this distinction further.

A.9 HUREF

HuRef derives model fingerprints from parameter-space invariants designed to remain stable under common post-training transformations (Zeng et al., 2024). Because its invariant construction is method-specific, WMMD-Bench retains the native invariant-comparison statistic rather than replacing it with a simpler parameter-similarity surrogate:

$$s_{\mathrm{HuRef}}(M) = \mathrm{ICS}(M, T),$$

where T denotes the registered reference model. The frozen decision rule is

$$d_{\mathrm{HuRef}}(M) = \mathbf{1}[s_{\mathrm{HuRef}}(M) > \tau_{\mathrm{HuRef}}], \qquad \tau_{\mathrm{HuRef}} = 4.119894027709961.$$

ICS is neither an ownership probability nor a retained-watermark fraction, and negative values are valid detector outputs. Same-lineage and cross-lineage measurements are therefore interpreted directly on this native scale.

A.10 AWM

AWM identifies model lineage from attention weight matrices using permutation-aware weight comparison and unbiased linear CKA (Zeng et al., 2026). Let $\mathcal{W}$ denote the fixed set of selected query and key attention-weight comparisons used by the benchmark. For reference weights W^T and candidate weights W^M, the native score is the mean

$$s_{\mathrm{AWM}}(M) = \frac{1}{|\mathcal{W}|} \sum_{W \in \mathcal{W}} \mathrm{CKA}_u\big(W^T, W^M\big)\,,$$

where CKA_u denotes the registered unbiased linear CKA calculation. The 3B detector decision is

$$d_{\mathrm{AWM}}(M) = \mathbf{1}[s_{\mathrm{AWM}}(M) > \tau_{\mathrm{AWM}}]\,, \qquad \tau_{\mathrm{AWM}} = 0.0017953364917795106.$$

The clean-Qwen score equals this strict calibrated threshold. High same-lineage scores remain compatible with pre-existing weight-space similarity, while cross-lineage threshold crossings retain the calibration limitation described in Appendix G.3. The representative 7B extension uses its own frozen calibration threshold (Appendix I).

A.11 ZeroPrint

ZeroPrint approximates input-sensitive gradient information from black-box responses using zeroth-order semantic perturbations and uses the resulting Jacobian-derived representation as a model fingerprint (Shao et al., 2026). Let z_T and z_M denote the fixed fingerprint representations extracted from the registered reference T and candidate model M. WMMD-Bench retains the native rescaled Pearson-correlation score

$$s_{\mathrm{ZP}}(M) = \mathrm{RescaledPearson}(z_T, z_M),$$

rather than redefining the method through a different response-similarity measure. Its frozen decision rule is

$$d_{\mathrm{ZP}}(M) = \mathbf{1}[s_{\mathrm{ZP}}(M) > \tau_{\mathrm{ZP}}]\,, \qquad \tau_{\mathrm{ZP}} = 0.6793505996465683.$$

The clean-Qwen score equals this strict threshold and is therefore negative. Because Qwen participates in calibration, threshold crossings of distilled Qwen students are interpreted together with the calibration analysis in Appendix G.3.

B Experimental Protocol

B.1 Models and Initialization

The main benchmark uses the following instruction-tuned checkpoints (Grattafiori et al., 2024; Qwen et al., 2024):

`meta-llama/Llama-3.2-3B-Instruct`
revision `0cb88a4f764b7a12671c53f0838cd831a0843b95`;
`Qwen/Qwen2.5-3B-Instruct`
revision `8f4992eda43eea7c770690ddc0de8f732da246f5`.

Every student starts independently from the corresponding unmodified instruction-tuned checkpoint, rather than from random weights or a previously distilled student. Bc does not continue from Ba, and XBb does not continue from XBa. Llama provides the same-lineage initialization, while Qwen provides the cross-lineage initialization.

For the same-lineage benchmark, the model family, model size, checkpoint revision, tokenizer, and chat template are fixed across matched student conditions. Students therefore begin from the same clean initialization, so differences after distillation are not confounded by changes in model capacity or pretrained starting point. Cross-lineage experiments deliberately replace the student family with Qwen while preserving the corresponding extraction and training protocol as closely as the tokenizer and architecture permit.

The active methods use their method-specific protected teachers. The five passive fingerprints use a shared unmodified reference model and a single trained student within each common distillation condition, which is then evaluated by all five passive detectors. Utility is likewise evaluated once for each shared passive student.

B.2 Distillation Datasets

Each method-specific source dataset contains 20,000 teacher-response records. A record contains an instruction, an optional input, the teacher response, and a sample identity. Active teachers produce method-specific response distributions, while the five passive fingerprints use a shared ordinary-response dataset. Matching the record budget therefore does not imply identical response content across methods, but it fixes the number of training records within the benchmark comparisons.

Within each matched direct/transformed comparison, the transformed-response dataset is constructed as a paired descendant of its corresponding parent. The instruction, input, sample identity, and record order are preserved, and only the response target may change. Direct-response training uses the original teacher response, whereas transformed-response training uses the paired transformed response, including cases that remain unchanged. This pairing keeps the extraction queries and sample composition fixed while varying the response supervision supplied to the student. Detailed transformation accounting is reported in Appendix C.4.

Historical-data boundary. The original historical PN-FP and SCW Ba datasets are unavailable; later follow-up experiments therefore use protocol-matched reconstructed parents. These reconstructions preserve the benchmark protocol but are not assumed to reproduce the exact historical samples.

B.3 Serialization and Response-Only Masking

User content consists of the instruction followed, when the input field is nonempty, by two newlines, `Input:`, a newline, and the input text. The corresponding tokenizer's chat template serializes the user turn and assistant answer. Prompt-only serialization includes the assistant generation header, whereas full serialization includes the answer. Encoding uses `add_special_tokens=False` to avoid duplicate automatic insertion of special tokens.

The prompt-token prefix and batch padding are excluded from the loss using the ignore index. Examples are capped at 1,024 tokens, and supervision is restricted to assistant-response positions. In XBa and XBb, teacher responses are reserialized using Qwen's own tokenizer and chat template rather than transferring Llama token identities. Each ownership detector retains its own verification interface, which need not coincide with the student-training serialization.

B.4 Optimization

We match the training budget and optimization protocol across distillation conditions whenever the objective permits. The main benchmark uses 20,000 training records, three epochs, an effective batch size of eight, learning rate 10^{-5}, BF16 precision, and seed 42. The protocol-matched follow-up and Bc runs additionally use the optimizer and scheduler settings summarized in Table A3. Response-only masking and the 1,024-token sequence limit are retained across these conditions.

Training updates the full student model without LoRA or quantized student weights. Under the common 20,000-record, three-epoch schedule with effective batch size eight, training comprises 7,500 optimizer steps. The Bc objective is specified in Appendix C.3.

CTCC uses its protected teacher with the frozen adapter while training a fresh full-parameter student. iSeal preserves the teacher-head treatment described in Appendix A.4. Final students are saved and freshly reloaded before endpoint ownership verification and utility evaluation.

B.5 Compute Environment

All reported experiments use a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU with BF16 model computation. The implementations use PyTorch and Transformers. The representative

Table A3: Optimization settings for Bc and the protocol-matched follow-up runs. The record budget, optimization horizon, effective batch size, learning rate, seed, and response-only supervision are matched across comparable conditions.

Setting	Value
Training records / epochs	20,000 / 3
Optimizer steps	7,500
Per-device batch / gradient accumulation	8 / 1
Effective batch size	8
Optimizer	PyTorch AdamW (`adamw_torch`)
Learning rate / weight decay	10^{-5} / 0
Warmup / schedule	225 steps / cosine decay
Maximum sequence length	1,024 tokens
Precision / trainable parameters	BF16 / full parameter set
Seed	42

7B extension uses the same GPU class; its method-specific software and optimizer settings are reported in Appendix I.1.

C DISTILLATION CONDITIONS

Ba, Bb, and Bc denote direct-response, transformed-response, and soft-logit distillation, respectively. XBa and XBb use the corresponding cross-lineage Qwen student. The superscript in Bb*/XBb* identifies the final transformed-response condition reported in the main results; unstarred Bb/XBb in the prose refer to the same condition and do not denote a separate supervision objective.

For an extraction query x_i, let y_i^T denote the response generated by the protected teacher and $\tilde{y}_i^T$ its semantic-preserving transformation. The three supervision settings differ primarily in the teacher information available to the student:

$$\mathcal{D}_{\mathrm{Ba}} = \{(x_i, y_i^T)\}_{i=1}^N, \qquad \mathcal{D}_{\mathrm{Bb}} = \{(x_i, \tilde{y}_i^T)\}_{i=1}^N,$$

while Bc additionally exposes the teacher predictive distribution at supervised response positions. All three conditions use $N = 20{,}000$ extraction records.

C.1 DIRECT-RESPONSE DISTILLATION

For Ba and XBa, ordinary extraction queries are answered by the method-specific protected teacher, or by the shared passive reference, to form a frozen 20,000-record response dataset. The student is trained on the resulting teacher responses using response-only supervised fine-tuning. Teacher responses are generated before student training and are not resampled across epochs.

Ba uses a fresh same-lineage Llama student, while XBa uses a fresh cross-lineage Qwen student. Both follow the serialization and masking rules in Appendix B.3.

C.2 SEMANTIC-PRESERVING RESPONSE TRANSFORMATION

The transformed-response condition modifies the teacher response before student training while preserving the extraction query and sample identity. We use Qwen2.5-3B-Instruct as an external transformation model and rewrite only the teacher answer. The transformation prompt requires preservation of the original meaning, factual content, entities, numbers, stance, qualifications, and answer function while allowing wording and sentence structure to change when a natural rewrite is possible.

The transformation is detector-agnostic: it receives no detector queries, registered watermark information, or watermark-specific optimization signal. It is related to pre-distillation paraphrasing (Pan et al., 2025), but explicitly retains responses that are too short or atomic to rewrite without changing their meaning.

Generation and fallback. Sampling uses temperature 0.7, top-p 0.9, and seed 42. Generation length is bounded relative to the source response, and each non-atomic response receives at most three transformation attempts. Outputs that are empty, malformed, contain transformation commentary, or leak prompt/control text are rejected and retried. If no acceptable rewrite is produced, the original response is retained. This procedure preserves the full 20,000-record training budget and allows atomic, identity, and fallback cases to remain in the transformed dataset.

Transformation prompt. The following system and user messages are used for the 3B answer transformation; the source answer replaces the indicated placeholder.

> **System.** You are a careful semantic-preserving transformation assistant. Rewrite only the supplied response using different wording and sentence structure when a natural rewrite is possible. Preserve exactly the same meaning, factual content, entities, numbers, stance, qualifications, and answer function. Do not add or remove information. Output only the final response. Do not explain or describe the rewriting process, mention these instructions, add labels or wrappers, or use watermark or detector information. If the response is too short, atomic, or cannot be meaningfully rewritten without changing its meaning, return the original response exactly as it is.
>
> **User.** Transform the response below under the rules above. Return only its final semantic-preserving form. For a response that is too short, atomic, or unsafe to rewrite without changing meaning, return it exactly unchanged.
>
> `<SOURCE_ANSWER>`
> `{source_answer}`
> `</SOURCE_ANSWER>`

The resulting responses are frozen before student training. Bb therefore evaluates the complete transformation procedure, including responses that remain unchanged, rather than assuming that every teacher answer is paraphrased.

C.3 Soft-Logit Distillation

Bc augments hard-target supervision with the teacher's predictive distribution. For a supervised response position t, let $z_T^{(t)}$ and $z_S^{(t)}$ denote the teacher and student logits, and define

$$p_T^{(\tau,t)} = \mathrm{softmax}\Big(z_T^{(t)}/\tau\Big), \qquad p_S^{(\tau,t)} = \mathrm{softmax}\Big(z_S^{(t)}/\tau\Big).$$

At the assistant-response positions $\mathcal{M}$, Bc uses the objective in Equation 4,

$$\mathcal{L}_{\mathrm{Bc}} = \frac{1}{|\mathcal{M}|}\sum_{t\in\mathcal{M}}\Big[0.5\,\mathrm{CE}\Big(y_t, p_S^{(1,t)}\Big) + 0.5\,\tau^2 D_{\mathrm{KL}}\Big(p_T^{(\tau,t)} \,\|\, p_S^{(\tau,t)}\Big)\Big], \qquad \tau = 2.$$

The teacher remains frozen and supplies the full-vocabulary distribution without gradient updates or top-k approximation. Prompt and padding positions are excluded from both loss terms. Bc starts from the clean Llama initialization and uses the same 20,000-record budget, 7,500-step horizon, and optimization schedule in Table A3.

We evaluate Bc endpoints for all five active watermarks and the shared passive student. Acquisition trajectories are not measured for Bc. Cross-lineage Bc is not evaluated because Llama and Qwen do not share aligned token identities or vocabulary dimensions; introducing a vocabulary mapping or sequence-level alternative would define a different distillation objective.

C.4 Transformation Statistics

Table A4 reports literal response changes and supervised token counts for the five active-method parent/descendant pairs. Table A5 separates rephrased, atomic, identity, and fallback cases, including the shared passive dataset. Every record is retained; an unchanged response remains part of the transformed-response condition rather than being treated as missing data.

Table A4: Response changes and supervised tokens per dataset pass. Every row contains 20,000 records. Token counts use the documented Llama training formatter and exclude masked prompt and padding positions; they are not cross-tokenizer token-budget estimates.

Method	Changed	Unchanged	Parent tokens	Bb* tokens
PN-FP	14,650	5,350	414,693	487,363
EverTracer	12,099	7,901	251,646	363,292
CTCC	14,504	5,496	334,104	431,810
iSeal	12,577	7,423	233,202	305,839
SCW	12,718	7,282	313,586	405,499

Table A5: Response-transformation processing modes. Atomic answers bypass rewriting; identity denotes an unchanged transformation-model output; fallback retains the source response after unsuccessful transformation attempts. Each row sums to 20,000 records.

Method	Rephrased	Atomic	Identity	Fallback
PN-FP	14,650	3,681	1,646	23
EverTracer	12,099	4,864	2,959	78
CTCC	14,504	2,956	2,304	236
iSeal	12,577	4,791	2,601	31
SCW	12,718	5,066	2,165	51
Passive shared	13,061	4,635	2,256	48

The paired active datasets preserve all 20,000 sample identities together with their corresponding instruction and input fields. Under the documented 1,024-token training formatter, the audited active datasets retain supervised response tokens without truncation. The shared passive transformation contains 13,061 changed and 6,939 unchanged responses, including 4,635 atomic cases. These statistics show that Bb is not a uniform rewrite of every teacher answer and that equal record budgets do not imply equal supervised-token exposure across Ba and Bb.

D Full Benchmark Results

D.1 Full Native Detector Results

Tables A6–A8 expand the condensed results in Table 1 using the recorded endpoint precision. Counts, AUC, p-values, and similarity scores retain their native units and are not converted to a common normalized ownership score. N/A denotes a non-applicable canonical measurement rather than a zero-valued detector output or an unsuccessful student-training run. The five passive fingerprints share the same trained student within each common distillation condition.

Table A6: Full active-watermark endpoints. PN-FP, CTCC, and iSeal report counts out of 1,024, 95, and 200 probes, respectively; EverTracer reports member-oriented AUC, and SCW reports a p-value. Values correspond to the conditions summarized in Table 1.

Condition	PN-FP	EverTracer	CTCC	iSeal	SCW
Teacher / reference	956	1.0	95	179	0.0
Clean Llama	1	0.4417	0	0	0.9199569225311279
Ba	98	0.4987	0	0	0.8440861701965332
Bb*	72	0.4757	0	0	0.9529914855957031
Bc	110	0.59	0	0	0.8819239735603333
Clean Qwen	43	N/A	0	N/A	0.7446491718292236
XBa	84	N/A	0	N/A	0.9158855080604553
XBb*	67	N/A	0	N/A	0.9841499328613281

Table A7: Full passive-fingerprint endpoints for LLMPrint bit accuracy, REEF centered linear CKA, and HuRef ICS. Detection uses the method-specific rules in Appendix A.

Condition	LLMPrint	REEF	HuRef
Teacher / reference	1.0	1.0	99.99999237060547
Clean Llama	1.0	1.0	99.99999237060547
Ba	0.82	0.9542220066305829	99.99898529052734
Bb*	0.785	0.9672861269249347	99.99873352050781
Bc	0.86	0.9859362800187441	99.99800872802734
Clean Qwen	0.68	0.4546738923165847	-8.356179237365723
XBa	0.68	0.4712346890040632	-8.41477108001709
XBb*	0.695	0.47347651724262	-8.33336067199707

Table A8: Full passive-fingerprint endpoints for AWM mean query/key-weight unbiased linear CKA and ZeroPrint rescaled Pearson correlation. Clean-Qwen values equal their respective strict thresholds and are therefore negative.

Condition	AWM	ZeroPrint
Teacher / reference	1.0	1.0
Clean Llama	1.0	1.0
Ba	0.9999974135841643	0.8399372696876526
Bb*	0.9999961065394538	0.9170220196247101
Bc	0.9999972975679806	0.927045077085495
Clean Qwen	0.0017953364917795106	0.6793505996465683
XBa	0.0018811310743071122	0.70185287296772
XBb*	0.0018529736315550899	0.7015291452407837

The endpoint table and the checkpoint trajectories serve different purposes: Table 1 reports the benchmark endpoints, whereas Appendix E follows selected active watermarks throughout training. Small differences between endpoint-context and trajectory-context measurements are therefore treated descriptively rather than as independent evidence of a supervision effect.

D.2 Auxiliary Detector Statistics

Table A9 reports additional Bc outputs that clarify the primary detector measurements. EverTracer retains the empirical ROC operating point computed from 100 members and 100 nonmembers. CTCC reports suppression and normal-query controls separately from its trigger set. iSeal reports registered and held-out BLEU statistics in addition to the primary success counts.

Table A9: Auxiliary Bc detector measurements. EverTracer's threshold is an empirical ROC operating point rather than a general ownership cutoff. CTCC reports trigger and negative-control counts separately; iSeal BLEU values are shown to six decimal places.

Method	Statistic	Value
EverTracer	TPR / FPR	0.16 / 0.05
EverTracer	ROC operating threshold	-0.001890856020991323
EverTracer	Members / nonmembers	100 / 100
CTCC	Trigger / suppression / normal	0/95; 0/100; 0/105
iSeal	Registered successes	0/200
iSeal	Registered mean BLEU	2.543812
iSeal	Registered median BLEU	2.264081
iSeal	Held-out successes	0/100
iSeal	Held-out mean BLEU	2.663422
iSeal	Held-out median BLEU	2.468146
SCW	p-value	0.8819239735603333
SCW	Queries / decision	1,000 / not detected

EverTracer's adapted Qwen AUC values are 0.523 for XBa and 0.5227 for XBb*. They are reported separately from the canonical N/A cells because the tokenizer change alters the underlying likelihood measurement (Appendix G.1).

D.3 Same-Lineage Utility

Utility evaluation uses ARC-Challenge normalized accuracy and TruthfulQA MC2 (Clark et al., 2018; Lin et al., 2022). Table A10 reports the available same-lineage clean/reference measurements, while Table A11 reports the Ba, Bb*, and Bc student endpoints. Because the historical evaluations were not all produced in a single matched clean-reference context, we report the recorded measurements directly rather than constructing a common same-lineage utility delta.

Table A10: Available same-lineage utility references. Values are rounded to six decimal places. The listed clean/reference measurements belong to their recorded evaluation contexts and are not pooled into a single common baseline.

Evaluation context	Model	ARC-Challenge	MC2
EverTracer context	Clean Llama	0.446246	0.505687
EverTracer context	Protected teacher	0.421502	0.514798
Passive shared context	Clean Llama	0.448805	0.505451

Table A11: Same-lineage utility endpoints on ARC-Challenge normalized accuracy and TruthfulQA MC2. Values are rounded to six decimal places where the recorded precision permits. Passive utility is measured once per shared student.

Student	Condition	ARC-Challenge	MC2
PN-FP	Ba	0.478669	0.463480
PN-FP	Bb*	0.490614	0.486138
PN-FP	Bc	0.478669	0.468550
EverTracer	Ba	0.505119	0.493715
EverTracer	Bb*	0.498294	0.475597
EverTracer	Bc	0.497440	0.500700
CTCC	Ba	0.483788	0.450614
CTCC	Bb*	0.489761	0.448507
CTCC	Bc	0.464164	0.468082
iSeal	Ba	0.481229	0.449818
iSeal	Bb*	0.489761	0.461524
iSeal	Bc	0.443686	0.456499
SCW	Ba	0.489761	0.474734
SCW	Bb*	0.504266	0.486187
SCW	Bc	0.494881	0.491111
Passive shared	Ba	0.497440	0.475541
Passive shared	Bb*	0.490614	0.454416
Passive shared	Bc	0.511092	0.500110

Across the reported same-lineage conditions, the students remain evaluable on both utility benchmarks. These measurements argue against a simple whole-model-collapse explanation for weak active-watermark retention, while remaining limited to the two evaluated tasks.

D.4 Cross-Lineage Utility

The matched clean-Qwen baseline is 0.4402730375426621 on ARC-Challenge and 0.5712170418936038 on TruthfulQA MC2. Table A12 reports all XBa and XBb* values together with their matched changes from clean Qwen. Deltas are computed as student minus clean Qwen before rounding under the same chat-template and BF16 evaluation setting. These utility measurements remain available even when a native ownership detector is non-applicable.

Table A12: Cross-lineage utility and matched changes from clean Qwen. Values and deltas are rounded to six decimal places. Passive fingerprints share one student per condition, and the two utility metrics remain separate.

Student	Condition	ARC	MC2	Δ ARC	Δ MC2
Clean Qwen	–	0.440273	0.571217	–	–
PN-FP	XBa	0.494881	0.501914	+0.054608	-0.069303
PN-FP	XBb*	0.468430	0.491763	+0.028157	-0.079454
EverTracer	XBa	0.516212	0.450975	+0.075939	-0.120242
EverTracer	XBb*	0.460751	0.469077	+0.020478	-0.102140
CTCC	XBa	0.468430	0.448092	+0.028157	-0.123125
CTCC	XBb*	0.439420	0.453441	-0.000853	-0.117776
iSeal	XBa	0.494881	0.429831	+0.054608	-0.141386
iSeal	XBb*	0.494027	0.467364	+0.053754	-0.103853
SCW	XBa	0.490614	0.491592	+0.050341	-0.079625
SCW	XBb*	0.497440	0.484784	+0.057167	-0.086434
Passive shared	XBa	0.514505	0.491376	+0.074232	-0.079841
Passive shared	XBb*	0.494881	0.512157	+0.054608	-0.059060

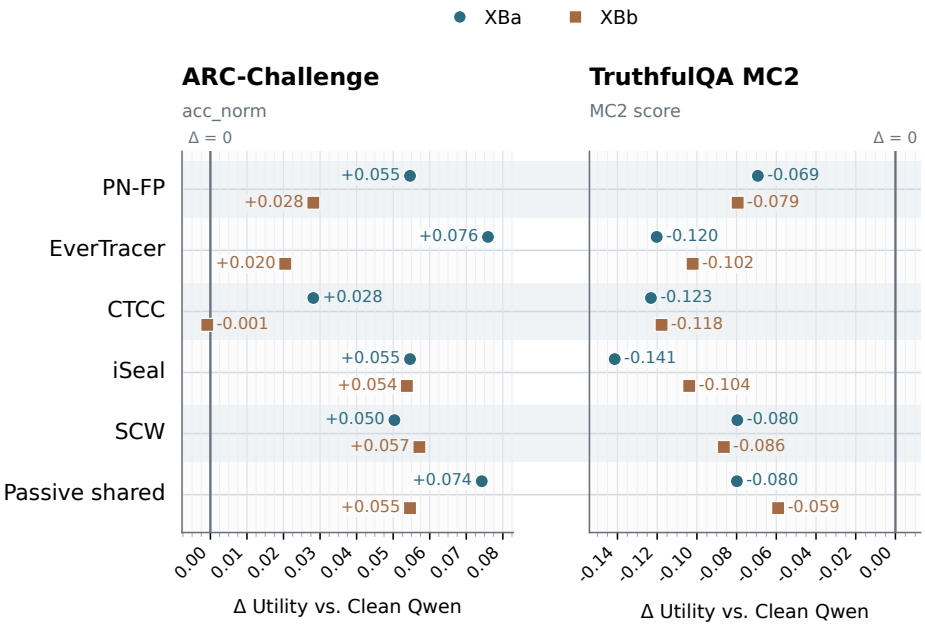

Figure A1: Cross-lineage utility changes for XBa and XBb*. Δ denotes student minus clean Qwen for ARC-Challenge normalized accuracy and TruthfulQA MC2. Labels show three decimal places, while point positions retain the recorded precision. Utility changes do not measure ownership-signal retention.

Cross-lineage utility changes are task-dependent rather than uniformly positive or negative. For example, the shared passive XBa student improves ARC-Challenge by 0.074232 while decreasing MC2 by 0.079841. This pattern is consistent with the students retaining nontrivial task performance despite weak active-watermark detection, but utility on these two benchmarks does not by itself establish preservation of general capability.

E Active-Watermark Acquisition Dynamics

This section reports all 55 measured checkpoint observations for the five active watermarks under same-lineage Ba. Each method follows one continuous training run with eleven sampled check-

points, rather than eleven independent seeds. Step zero is the measured clean Llama student, while the protected teacher is reported separately as a reference. Training loss is included only where recorded. At checkpoint 25, the nearest retained loss observation is from step 20; at the remaining nonzero checkpoints, loss is reported at the stated step. A dash at step zero denotes an unavailable training-loss measurement.

Across the sampled checkpoints, the observed trajectories are more consistent with limited acquisition than with strong acquisition followed by later forgetting. This interpretation is restricted to same-lineage Ba and to the measured checkpoints; behavior between checkpoints and acquisition dynamics under Bb, Bc, XBa, or XBb are not observed.

E.1 PN-FP

The PN-FP follow-up trajectory uses the protocol-matched parent described in Appendix B.2. Its endpoint is reported separately from the historical Ba measurement in Table 1.

Table A13: PN-FP acquisition under same-lineage Ba. Rates are exact fractions of 1,024 registered probes expressed as decimals. The protected teacher produces 956 matches.

Step	Matches / 1,024	Match rate	Loss
0	1	0.0009765625	–
25	1	0.0009765625	1.6851
50	3	0.0029296875	1.3432
100	58	0.056640625	0.6902
250	79	0.0771484375	0.6295
500	86	0.083984375	0.5212
1000	85	0.0830078125	0.5102
2000	94	0.091796875	0.4791
4000	99	0.0966796875	0.3148
6000	99	0.0966796875	0.3396
7500	99	0.0966796875	0.2766

PN-FP shows the clearest partial acquisition among the five active watermarks. The count rises from 1 match at initialization to 3 at step 50, 58 at step 100, 79 at step 250, and 99 by step 4,000. It remains at 99 through the final two measured checkpoints, approximately one tenth of the teacher's 956 matches. The small decrease from 86 to 85 between steps 500 and 1,000 is minor relative to the overall increase. The sampled trajectory therefore supports partial acquisition followed by a low plateau, rather than a period of strong acquisition followed by substantial collapse.

E.2 EverTracer

Table A14: EverTracer acquisition under same-lineage Ba with the fixed Llama-tokenizer interface. Each checkpoint uses 100 members and 100 nonmembers. TPR and FPR refer to the empirical operating point with FPR ≤ 0.05; thresholds are rounded to nine decimal places.

Step	AUC	TPR	FPR	ROC threshold	Loss
0	0.4417	0.05	0.05	-0.002477288	–
25	0.4401	0.07	0.05	-0.002617531	2.7436
50	0.4385	0.07	0.05	-0.002677994	1.8636
100	0.4457	0.07	0.05	-0.002651489	1.1200
250	0.443	0.06	0.05	-0.002692191	0.8426
500	0.4618	0.08	0.05	-0.003011572	0.7194
1000	0.4768	0.07	0.05	-0.002763159	0.6446
2000	0.5027	0.09	0.05	-0.003184851	0.8335
4000	0.4999	0.06	0.05	-0.002804719	0.4623
6000	0.5009	0.05	0.05	-0.002497874	0.3361
7500	0.5016	0.05	0.05	-0.002543277	0.3625

EverTracer remains near chance throughout the sampled trajectory. AUC ranges from 0.4385 to 0.5027 and ends at 0.5016, compared with the protected teacher's AUC of 1.0. The member-oriented score direction is fixed throughout, and changes in the empirical operating threshold arise from the same ROC procedure rather than from changes to the verifier. No sampled checkpoint shows a high-AUC acquisition phase followed by subsequent collapse.

E.3 CTCC

Table A15: CTCC acquisition under same-lineage Ba. Trigger, suppression, and normal columns report exact target activations in disjoint query sets.

Step	Trigger / 95	Suppress / 100	Normal / 105	Loss
0	0	0	0	–
25	0	0	0	2.7513
50	0	0	0	1.9597
100	0	0	0	1.0425
250	0	0	0	0.9549
500	0	0	0	0.7597
1000	0	0	0	0.7892
2000	0	0	0	0.6824
4000	0	0	0	0.4092
6000	0	0	0	0.4400
7500	0	0	0	0.6513

CTCC shows no registered exact-response acquisition at any sampled checkpoint. All 95 trigger counts remain zero, and the 100 suppression and 105 normal-query controls also remain at zero, while the protected teacher produces 95/95 trigger activations under the same operational rule. This result concerns the registered exact target and does not measure semantic variants or other unregistered cross-turn behaviors.

E.4 iSeal

Table A16: iSeal acquisition under same-lineage Ba. Registered and held-out success counts use 200 and 100 queries, respectively. Mean and median BLEU are reported for the registered set and rounded to six decimal places.

Step	Reg. / 200	Held-out / 100	Mean BLEU	Median BLEU	Loss
0	0	0	2.332514	2.009208	–
25	0	0	2.355856	2.069724	2.6178
50	0	0	2.354638	2.037041	1.6426
100	0	0	2.309473	1.985097	1.0254
250	0	0	2.455193	2.174992	0.7543
500	0	0	2.601366	2.312137	0.6759
1000	0	0	2.600376	2.266927	0.7295
2000	0	0	2.595570	2.251498	0.7429
4000	0	0	2.589713	2.273074	0.4458
6000	0	0	2.581606	2.289404	0.3516
7500	0	0	2.589418	2.260249	0.3609

iSeal produces no registered or held-out successes at any sampled checkpoint. Registered mean BLEU remains between approximately 2.31 and 2.60, far below the per-query success criterion of 50. The auxiliary overlap statistics therefore provide no evidence of a successful acquisition peak hidden by the zero success counts. The protected teacher, by comparison, produces 179/200 registered successes.

E.5 SCW

Table A17: SCW acquisition under same-lineage Ba. Every checkpoint uses 1,000 queries and the fixed decision threshold $\alpha = 0.001$. p-values retain the recorded precision.

Step	p-value	Detected	Loss
0	0.5813101530075073	no	–
25	0.8303400874137878	no	1.5583
50	0.6046135425567627	no	1.3581
100	0.8523297309875488	no	0.5416
250	0.8244985342025757	no	0.4974
500	0.8764297962188721	no	0.4385
1000	0.6776783466339111	no	0.5142
2000	0.5024667382240295	no	0.3964
4000	0.949506402015686	no	0.2882
6000	0.6830558776855469	no	0.2794
7500	0.895175039768219	no	0.2649

SCW remains undetected at every sampled checkpoint. Its p-values fluctuate within the non-significant region and never cross the fixed $\alpha = 0.001$ decision threshold. Because non-significant p-value differences are not calibrated as a continuous measure of watermark strength, these fluctuations do not indicate a hidden acquisition-and-collapse pattern. The trajectory is therefore interpreted only through its registered detection rule, under which no sampled student checkpoint is positive.

F Watermark-Signal Exposure Analysis

The signal-exposure analysis tests whether registered verification information appears exactly in the available distillation data. We audit the five 20,000-record active-watermark parent datasets used in the corresponding follow-up experiments together with their paired transformed-response descendants. The audit is deliberately focused on direct registered exposure: it asks whether the registered verification content itself appears in the training records under each method-specific matching rule.

Table A18: Exact registered overlap in the active-watermark distillation datasets. Parent and Bb* denote the paired direct-response source dataset and its transformed-response descendant. All counts are out of 20,000 records. The matching rule follows the registered verification structure of each method.

Method	Registered information and matching rule	Parent	Bb*
PN-FP	Key substrings in training questions; registered target substrings in answers and retained supervised labels	0	0
EverTracer	Exact whole-field matches to any of the 200 registered verification documents in the training fields	0	0
CTCC	Exact trigger-set questions; `IAMALIVE` target occurrences in answers and retained supervised labels	0	0
iSeal	Exact whole-field matches to any of the 200 registered plaintexts in instruction, input, answer, or combined question fields	0	0
SCW	Exact occurrences of any of the 1,000 registered French evaluation instructions	0	0

Table A18 shows zero exact registered overlap under all five method-specific checks. This result rules out direct exposure to the registered verification examples under the stated matching rules, but it does not imply that the distillation data contain no information related to the ownership signals. The student may still receive shared tokens, recurring lexical patterns, semantically related content, broader teacher-response regularities, distributional information, or representation-level support that partially overlaps with the behavior required by a watermark.

The scope of the exact-match audit also differs across mechanisms. EverTracer's check covers whole-field matches to its 200 registered verification documents rather than the complete fingerprint-training corpus. iSeal plaintext matching does not reproduce its keyed embedding interface. Likewise, SCW's zero overlap with the registered French instructions does not exclude other French-language or semantically related supervision. The reported zeros should therefore be interpreted as absence of direct registered example overlap, rather than absence of all watermark-relevant information in the distillation data.

This distinction is important for interpreting the acquisition trajectories. PN-FP develops partial registered matching behavior despite zero exact registered key–target overlap in the audited training data (Appendix E.1). Exact exposure to the registered key–target pairs is therefore not necessary for the partial PN-FP acquisition observed under this protocol. CTCC, in contrast, shows no registered exact-response acquisition at the sampled checkpoints even though its exact registered-trigger frequency is likewise zero. The contrast indicates that exact registered exposure alone is insufficient to explain the different acquisition outcomes and is consistent with the possibility that ordinary distillation supervision provides different degrees of indirect support for different watermark structures.

G Cross-Lineage Detector Applicability

A detector can return a numerical value without preserving the scientific meaning of its original measurement. Cross-lineage applicability therefore concerns whether the verification interface, statistic, and decision rule remain comparable after changing the student model family, separately from whether student training or utility evaluation succeeds.

G.1 Cross-Tokenizer Applicability

EverTracer combines fixed original and perturbed texts with a token window, mean next-token likelihood, and member-oriented aggregation. The portability analysis uses 200 original texts and 2,000 perturbations. Replacing the Llama tokenizer with Qwen changes the text retained under the 128-token limit for 1,477 of the 2,200 items. Even when the Llama-visible text is preserved first, 1,262 items still exceed Qwen's 128-token window.

For a token sequence $z_1, \ldots, z_n$, the likelihood primitive depends on

$$-\frac{1}{n-1}\sum_{t=2}^{n}\log p(z_t \mid z_{<t}),$$

and EverTracer subsequently applies the exponential transformation described in Appendix A.2. A tokenizer change alters both the prediction units and the number of prediction events; truncation can additionally change the visible text. Consequently, applying the same likelihood formula after switching tokenizers does not preserve the original measurement semantics.

The adapted Qwen evaluation preserves the registered texts, perturbations, labels, member direction, and ROC/FPR procedure, while replacing the tokenizer and retaining the 128-token limit. It yields AUC 0.523 for XBa and 0.5227 for XBb*. We report these values as exploratory measurements rather than as canonical EverTracer results, and retain N/A in the canonical cross-lineage cells.

G.2 Detector Portability

We distinguish three reporting categories. A *canonical* result preserves the benchmark's defined probe semantics, native statistic, and decision rule. An *exploratory* result applies an adapted calculation whose measurement semantics are not established as equivalent to the canonical verifier. A *non-applicable* result is reported as N/A when preserving the registered verification interface would require redefining the method. These categories describe measurement validity rather than watermark strength.

PN-FP, CTCC, and SCW retain operational cross-lineage verification interfaces. PN-FP uses the same registered text keys and targets. CTCC preserves the same registered probes and the exact, case-sensitive target rule after whitespace stripping. SCW preserves the query set, statistical procedure, and decision direction. For CTCC, "canonical" refers to the fixed WMMD-Bench operational

adaptation defined in Appendix A.3, rather than to exact equivalence with every component of the original verifier.

EverTracer is reported as exploratory across tokenizers for the reasons above. iSeal is non-applicable under the evaluated Qwen architecture because its keyed embedding interface is architecture-dependent. Introducing a new projection, changing the registered dimensions, or re-registering the keyed interface for Qwen would alter the ownership mechanism being evaluated. Canonical cross-lineage iSeal detection is therefore reported as N/A. These portability constraints affect ownership verification rather than the existence of the corresponding Qwen students or their utility measurements.

G.3 Calibration and Clean-Control Interpretation

All five passive fingerprints are already positive on the clean same-lineage Llama student before distillation. Their positive Ba, Bb*, and Bc results therefore cannot by themselves establish that fingerprint information was newly transferred from the protected teacher. High post-distillation similarity remains compatible with pre-existing model lineage and initialization.

Cross-lineage threshold interpretation requires a separate control. For REEF, AWM, and ZeroPrint, clean Qwen participates in threshold calibration. Its score equals the corresponding strict threshold and is therefore classified as negative. Distilled Qwen students exceed these thresholds, but a model used to define the threshold is not an independent held-out negative for validating the same rule. These threshold crossings remain valid detector measurements, but they do not provide independent evidence of distillation-induced fingerprint transfer or population-level false-positive rates. LLMPrint and HuRef remain below their respective thresholds on the evaluated Qwen students.

Even when a detector remains operational across model families, replacing Llama with Qwen changes several model properties simultaneously, including architecture, tokenization, pretraining, and representation geometry. The cross-lineage contrast therefore shows that passive detectability is strongly associated with model relatedness under the evaluated conditions, but it does not isolate any single model property as the source of that difference.

H Cross-Dimensional Analysis and Design Implications

The preceding sections report the benchmark results along separate dimensions: ownership mechanism, distillation supervision, student lineage, acquisition dynamics, signal exposure, and detector applicability. We next consider these dimensions jointly. The goal is not to introduce a new aggregate ownership score, but to identify patterns that are only visible when the native measurements, clean controls, trajectories, and applicability analyses are interpreted together.

H.1 Active and Passive Ownership Signals Behave Differently

The strongest cross-method contrast is between active watermarks and passive fingerprints. Across the five active watermarks, distillation substantially weakens the registered ownership evidence. PN-FP retains only a small fraction of the teacher's registered signatures, EverTracer remains near chance under direct-response distillation, CTCC and iSeal produce no registered successes at the reported student endpoints, and SCW remains above its detection threshold. The checkpoint trajectories in Appendix E further show that these weak endpoints are not generally preceded by a strong acquisition phase. PN-FP exhibits partial acquisition followed by a low plateau, while the other four active methods show no comparable registered acquisition peak at the sampled checkpoints.

Passive fingerprints exhibit the opposite endpoint pattern. All five remain positive under the same-lineage Ba, Bb*, and Bc conditions, often with scores close to their reference regions. However, all five are already positive on the clean same-lineage Llama student before distillation. Their apparent stability therefore has a different interpretation from that of an active watermark: high post-distillation similarity does not by itself show that the student acquired new ownership information from the protected teacher.

Taken together, the two method families expose different limitations. For the evaluated active watermarks, the main difficulty is acquisition: the registered signal is weakly expressed or absent in the

distilled student. For the passive fingerprints, the main difficulty is specificity: strong same-lineage detectability can already be present before teacher supervision. Robust ownership verification under model extraction therefore requires both a signal that can survive knowledge transfer and evidence that remains specific to the protected model rather than to a broader model lineage.

H.2 Changing the Supervision Channel Does Not Remove the Active–Passive Asymmetry

The three same-lineage distillation settings expose different amounts and forms of teacher information. Ba provides ordinary teacher responses, Bb* changes their surface realization through a detector-agnostic semantic-preserving transformation, and Bc additionally exposes the teacher's predictive distribution at supervised response positions (Appendix C).

For the active watermarks, changing the supervision channel affects some native measurements but does not change the overall pattern. PN-FP and EverTracer are stronger under Bc than under Ba, indicating that richer teacher supervision can increase some ownership-related measurements. The recovery is nevertheless limited and method-dependent: CTCC and iSeal remain at zero registered successes, and SCW remains undetected. The transformed-response condition is similarly non-uniform. PN-FP decreases under Bb*, whereas the other active methods do not show a common attenuation pattern. Thus, neither altered response surface form nor additional distribution-level information reliably restores the evaluated active ownership signals.

The passive fingerprints again behave differently. Their same-lineage scores remain positive across Ba, Bb*, and Bc, but the clean-Llama controls show that this persistence cannot be attributed solely to the amount of teacher information transferred during distillation. More informative supervision is therefore not sufficient, by itself, to explain ownership robustness across either method family.

These results suggest that watermark survival depends on more than the quantity of teacher information available to the student. The transferred information must also support the behavior or model property used by the verifier. This distinction is consistent with the method-specific recovery observed under Bc and with the different verification structures summarized in Appendix A.6.

H.3 Model Lineage Strongly Shapes Passive Fingerprint Detectability

The same-lineage and cross-lineage comparisons provide a second separation between active and passive mechanisms. For the active methods with applicable cross-lineage verification, replacing the Llama student with Qwen does not produce a general recovery of the registered signals. PN-FP increases above its clean-Qwen baseline after cross-lineage distillation, but remains far below the protected-teacher count. CTCC remains at zero registered activations, and SCW remains undetected. The active-watermark conclusion is therefore not restricted to a single same-lineage student setting.

Passive fingerprints show a substantially stronger dependence on model relatedness. Their same-lineage clean and distilled scores lie close to the reference regions, whereas Qwen occupies markedly different native-score regions. This contrast is consistent across behavioral, representation-based, and weight-based passive fingerprints, although the magnitude differs by method.

The cross-lineage results must be interpreted together with Appendix G. Replacing Llama with Qwen simultaneously changes architecture, tokenizer, pretraining, and representation geometry, and several passive thresholds use Qwen during calibration. Consequently, the cross-lineage comparison does not isolate a single causal property. It nevertheless provides a consistent empirical pattern: passive fingerprint detectability is strongly associated with model relatedness, while high same-lineage post-distillation scores do not constitute clean evidence of newly transferred fingerprint information.

This distinction clarifies why active and passive robustness should not be interpreted using the same notion of "retention." An active watermark is designed to introduce a registered signal that should remain observable after extraction. A passive fingerprint may instead remain detectable because the student already occupies a similar behavioral, representational, or parameter-space region. Clean initialization is therefore an essential reference for passive methods, rather than an auxiliary control.

H.4 Signal Structure Provides Context for the Acquisition Results

The five active watermarks also differ in the structure required for successful verification. Appendix A.6 summarizes this distinction: PN-FP is evaluated through individual registered key–target associations, EverTracer through document-level likelihood evidence, CTCC through a multi-turn contextual relation, iSeal through a keyed verification interface, and SCW through aggregate statistical evidence conditioned on semantic content.

The trajectories and exposure analysis provide a useful contrast between these structures. PN-FP is the only evaluated active method that shows clear partial acquisition along the sampled Ba trajectory. CTCC, by comparison, produces no registered exact-response activations at any measured checkpoint. Yet the exact-exposure audit in Appendix F finds no registered key–target exposure for PN-FP and no registered trigger exposure for CTCC in the audited distillation data.

The PN-FP result therefore shows that direct appearance of the registered key–target pairs is not necessary for partial acquisition under the evaluated protocol. At the same time, the zero-overlap audit does not imply that the distillation data contain no watermark-relevant information. As discussed in Appendix F, ordinary teacher responses can still share tokens, lexical patterns, semantic content, response regularities, distributional information, or representation-level structure with the behavior used for verification.

The different PN-FP and CTCC trajectories are therefore consistent with the possibility that ordinary response supervision provides different degrees of indirect support for different ownership-signal structures. In particular, PN-FP verification depends on an individual query–response association, whereas CTCC requires the relevant contextual relation across dialogue turns before its registered target behavior can be expressed. We treat this as an interpretation of the observed contrast rather than as a separate measured complexity variable. The benchmark evidence establishes the difference in acquisition behavior and the absence of exact registered exposure; the structural comparison provides a way to interpret these observations jointly.

H.5 Implications for Distillation-Resilient Watermark Design

The combined results suggest a practical tension for ownership-mechanism design. A signal that is highly specialized to a rare verification interaction may be easy to distinguish in the protected teacher but weakly supported by the ordinary responses used for extraction. At the other extreme, a signal tied primarily to broad model geometry or lineage may remain detectable after distillation while providing limited evidence that information was transferred from a particular protected teacher.

Response-based extraction in WMMD-Bench trains the student on ordinary question-answering and instruction-following records rather than on the registered ownership queries themselves. Under such supervision, behaviors that recur naturally in ordinary responses are more likely to be repeatedly exposed to the student than behaviors requiring a specialized verification context. The acquisition pattern observed for PN-FP, together with the absence of registered examples in the exposure audit, is consistent with ownership information being supported indirectly rather than only through exact trigger repetition.

This observation motivates a broader design principle. Distillation-resilient active watermarks may benefit from ownership signals that are expressed through subtle, recurring linguistic or semantic preferences that naturally appear during ordinary model use. Examples include consistent lexical choices, discourse-marker preferences, syntactic phrasing, response organization, or other low-level behavioral regularities. Such signals need not rely on a single visible phrase; the relevant property is that the ownership behavior can recur within ordinary responses without requiring a rare or highly structured interaction.

Transferability alone, however, is insufficient. The passive-fingerprint results show that a signal can remain highly detectable across related models while providing weak evidence of transfer from one protected teacher. Accordingly, a useful distillation-resilient ownership mechanism should jointly satisfy three requirements: it should be expressed often enough to be acquired through realistic response supervision, remain sufficiently specific to the protected model to avoid clean-model positives, and preserve a meaningful verification interface across plausible student architectures and tokenizers.

These requirements connect the main empirical findings of WMMD-Bench. The active-watermark results emphasize acquisition, the passive-fingerprint results emphasize specificity, and the cross-lineage applicability analysis emphasizes verifier portability. Designing ownership signals that balance these three properties provides a direct interpretation of the benchmark results and a concrete direction for ownership mechanisms intended to remain reliable after model distillation.

I Representative 7B Extension

I.1 Experimental Setup

The representative extension evaluates PN-FP and AWM on the following checkpoint, with 6,738,415,616 parameters:

`meta-llama/Llama-2-7b-chat-hf`
revision `f5db02db724555f92da89c216ac04704f23d4590`.

Each student independently starts from the clean canonical model. PN-FP uses its selected protected teacher; AWM uses the unmodified reference. Their teacher-response datasets remain method-specific.

Each condition uses 20,000 records, three epochs, 7,500 optimizer steps, effective batch size eight, a 1,024-token maximum length, learning rate 10^{-5}, 225 warmup steps, cosine decay, zero weight decay, and seed 42. Training updates all parameters in BF16, with activation checkpointing, the native chat template, and assistant-response masking. Unlike the main 3B protocol, the optimizer is 8-bit AdamW (`bitsandbytes` 0.50.0). This reduces optimizer-state memory and is not numerically equivalent to the PyTorch AdamW implementation used at 3B; student weights remain BF16. The recorded training environment uses PyTorch 2.8.0 with CUDA 12.8 and Transformers 4.44.2; AWM verification uses a separate Transformers 4.51.3 environment on the same GPU model as the main benchmark.

Ba, Bb, and Bc retain their direct-response, transformed-response, and soft-logit meanings. The 7B transformation uses Qwen2.5-3B-Instruct with temperature 0.7, top-p 0.9, seed 42, and a source-length-dependent generation budget with minimum 64, multiplier 1.5, and cap 1,536. Its prompt is fixed for the 7B experiments. Soft-logit distillation uses the full-vocabulary objective in Equation 4, aligned shifted response positions, temperature 2, and equal CE/KL weights. In this extension, selected teacher logits are cast to BF16 and cached before student training, rather than recomputed online. No top-k approximation is used.

I.2 Ownership Results

Table A19: Representative 7B ownership results. PN-FP match rate is the count divided by 1,024; teacher-relative retention is the count divided by 804. AWM retains its native similarity scale and strict threshold 0.007566815861277831.

Method	Condition	Native score	Match rate (%)	Retention (%)
PN-FP	Teacher	804/1,024	78.515625	100.000000
PN-FP	Ba	53/1,024	5.175781	6.592040
PN-FP	Bb*	71/1,024	6.933594	8.830846
PN-FP	Bc	144/1,024	14.062500	17.910448
AWM	Reference	1.0000000000000000	–	–
AWM	Ba	0.9999952428042889	–	–
AWM	Bb*	0.9999933559447527	–	–
AWM	Bc	0.9999957373365760	–	–

PN-FP decreases from 804 teacher matches to 53 after Ba, 71 after Bb, and 144 after Bc. The largest student count is 17.910448% of the teacher count. This is a descriptive single-seed comparison and does not establish a statistically reliable supervision effect. AWM remains close to its reference similarity of 1.0 in all three conditions. Its threshold is fixed before student training as the maximum

score among three negative reference models: Gemma-2B, Qwen2.5-3B-Instruct, and Phi-3-mini-128k-instruct. Ties are negative. The high same-lineage student scores remain compatible with pre-existing weight-space similarity and do not identify newly acquired ownership information.

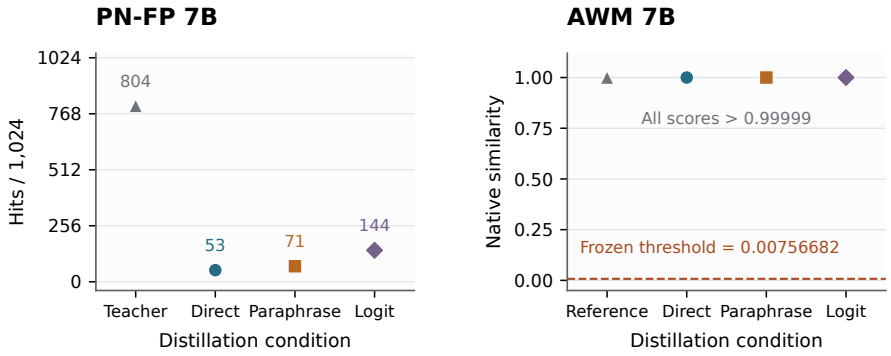

Figure A2: Representative 7B endpoints in separate native units. Direct, Paraphrase, and Logit denote Ba, Bb, and Bc, respectively. PN-FP shows registered matches, while AWM shows reference and student similarity with its frozen threshold. The panels do not share a normalized ownership score.

I.3 Utility Results

Table A20: 7B utility on ARC-Challenge normalized accuracy (1,172 examples) and TruthfulQA MC2 (817 examples). Percentages are rounded to six decimal places. PN-FP includes its selected protected teacher, while AWM includes the clean reference.

Method	Condition	ARC-Challenge (%)	TruthfulQA MC2 (%)
PN-FP	Teacher	42.064846	44.199762
PN-FP	Ba	46.843003	42.404140
PN-FP	Bb*	48.464164	43.801800
PN-FP	Bc	44.624573	42.416776
AWM	Clean/reference	36.604096	45.754716
AWM	Ba	46.843003	42.953599
AWM	Bb*	47.781570	43.869722
AWM	Bc	44.112628	42.639766

The clean Llama reference scores 36.604096% on ARC-Challenge and 45.754716% on MC2. All reported students have higher ARC-Challenge scores and lower MC2 scores than this reference. The direction therefore depends on the task, and two benchmarks do not establish preservation of general capability. The selected PN-FP teacher is reported separately from the clean reference.

I.4 Scope and Limitations

This is a representative robustness extension, not a controlled scaling experiment. The main benchmark uses Llama-3.2-3B-Instruct, whereas the extension uses Llama-2-7b-chat-hf and a different optimizer implementation; the logit-delivery implementation also differs. Parameter count cannot be isolated from these changes. Only two methods and one seed are evaluated. No 7B cross-lineage result or student acquisition trajectory is reported. The observations broaden the evidence for PN-FP and AWM under the evaluated conditions, without establishing a general scaling law or outcomes for the other eight mechanisms.

J Comparison with Prior Benchmarks and Related Evaluation Settings

J.1 Evaluation Scope

Table A21 compares the evaluation questions addressed by related work. It distinguishes robustness of generated text from ownership verification of a trained model, and changes to a text or model from training an extracted student. The table summarizes the cited studies; it is not a claim that their tools cannot support other experiments. We omit method-count rankings because algorithms, watermark families, and ownership mechanisms are different comparison units.

Table A21: Prior benchmarks and related evaluation settings. Each row describes the focus of the cited work rather than marking unverified absences. Model-level fingerprints and text watermarks can provide complementary provenance evidence while using different verification interfaces.

Work	Object of evaluation	Evaluated setting and evidence
WaterBench (Tu et al., 2024)	Generated-text watermarks	Joint evaluation of generation quality and detection across tasks and watermark strengths.
Mark My Words (Piet et al., 2025)	Generated-text watermarks	Quality, detection text length, and robustness to text perturbations.
MarkLLM (Pan et al., 2024)	Text-watermark implementations	Common interfaces, visualization, and evaluation of detectability, robustness, and text quality.
UWBench (Wu et al., 2026)	Unbiased text watermarks	Distributional effects, detectability, and robustness under token edits and paraphrasing.
UPKD (Pan et al., 2025)	Inherited text-watermark signals	Response-based student training, including cross-family extraction; pre-distillation paraphrasing and student inference-time neutralization.
PN-FP (Nasery et al., 2025)	Active model fingerprints	Scalable registered behavioral signatures and robustness to subsequent model modifications.
REEF (Zhang et al., 2025)	Passive model fingerprints	Representation similarity for identifying related models under model modifications.
EverTracer (Xu et al., 2025a)	Active model fingerprints	Probabilistic verification and robustness of registered likelihood-based evidence.
ADFP (Xu et al., 2026)	Fingerprints designed for extraction	Response-based distillation to Qwen and Llama students; fingerprinted and unfingerprinted training controls and empirical detection evaluation.
WMMD-Bench	Active and passive model ownership	Direct-response, transformed-response, and same-lineage soft-logit supervision; clean controls, sampled acquisition dynamics, and cross-lineage detector applicability.

J.2 Distillation, Controls, and Measurement Validity

Watermark learnability and radioactivity establish related settings in which output-distribution biases can be learned through training (Gu et al., 2024a; Sander et al., 2024). UPKD studies response-based unauthorized distillation, including a GLM teacher and Llama student, and evaluates both pre-distillation paraphrasing and inference-time watermark neutralization (Pan et al., 2025). Its teacher-access setting excludes teacher logits. Changing student decoding at inference time is therefore distinct from providing online teacher distributions during Bc training. Its unwatermarked-output training controls and data-size comparisons answer different questions from measuring a clean instruction-tuned student and checkpoints along one continuous training run.

ADFP explicitly designs output fingerprints for acquisition by extracted students, including Qwen and Llama students and controls trained on unfingerprinted teacher outputs (Xu et al., 2026). Its empirical ROC evaluation provides evidence about distinguishing those training conditions. Such

controls should not be conflated with the pre-distillation clean controls used here, just as a comparison of different training-data sizes should not be read as an acquisition trajectory from a single run.

The model-level mechanisms represented by PN-FP, REEF, and EverTracer provide complementary behavioral, representation, and likelihood interfaces (Nasery et al., 2025; Xu et al., 2025a; Zhang et al., 2025). WMMD-Bench evaluates these together with CTCC, iSeal, SCW, LLMPrint, HuRef, AWM, and ZeroPrint, retaining the benchmark definitions in Appendix A. Its contribution is the joint evaluation of heterogeneous mechanisms across supervision conditions, clean initialization, sampled acquisition dynamics, and student lineage. Detector portability is treated explicitly because a change in tokenizer or keyed interface can change the meaning of a score. The resulting evidence remains conditional on the selected models, datasets, and native decision rules; it does not establish superiority over every related evaluation setting.

K Reproducibility Notes

K.1 Supplementary Materials

The supplementary package includes the complete manuscript project, figure- and table-generation code, numerical endpoint and utility summaries, all 55 checkpoint observations, protocol configurations, detector thresholds, model revision information, and response-transformation specifications. These materials support inspection and reanalysis of the reported results. Content hashes are included in the supplementary artifact for provenance tracking. The training seed is 42; method-specific detector rules, model identities, and optimization settings are specified in Appendices A–C, with the separate 7B configuration in Appendix I.

K.2 Reproduction Boundaries

The package does not redistribute model weights, complete training datasets, all registered detector assets, complete raw experiment outputs, or a fully versioned execution environment. Repeating training and detector execution therefore requires additional model and data assets, subject to their applicable access and licensing terms; the included analysis code is not a self-contained release of every training and verification implementation. Scalar AUC summaries alone do not supply the per-example scores and labels needed to recompute ROC curves.

The unavailable historical PN-FP/SCW parents cannot be exactly reproduced from their reconstructed follow-ups, and clean base-model revisions do not identify every method-specific protected checkpoint. Historical same-lineage utility contexts also cannot be assumed to share a matched clean baseline. These boundaries are retained in the result presentation. All endpoint and trajectory comparisons are descriptive at the measured seed; the supplementary material does not provide repeated-seed uncertainty estimates or observations between sampled checkpoints.