FUNCTIONAL SUBSPACE WATERMARKING FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-side watermarking protects the ownership of large language models (LLMs) by leveraging model weights or internal representations. However, these internal signals often undergo complex distortions during practical model modifications, such as fine-tuning, quantization, pruning, and knowledge distillation, making reliable watermark extraction extremely challenging. Although model-side watermarking has been widely studied, existing methods still lack sufficient robustness against parameter-level and representation-space perturbations. To address this issue, we propose Functional Subspace Watermarking (**`FSW`**), which embeds multi-bit watermark information into a low-dimensional functional backbone of LLM internal representations. Specifically, **`FSW`** solves a generalized eigenvalue problem to jointly model functional sensitivity and compression invariance, thereby extracting watermark directions that are both task-critical and stable under common model modifications. We further introduce an adaptive spectral truncation strategy to select an appropriate spectral band that balances robustness and model utility, while incorporating a vector consistency constraint to mitigate representation drift during watermark injection. Experiments show that **`FSW`** achieves 100% bit recovery accuracy across diverse LLM backbones and, while maintaining competitive utility on most evaluated backbones, preserves watermark detectability under LoRA fine-tuning, noise perturbation, pruning, quantization, and backbone-preserving distillation. In addition, we further evaluate FSW under untargeted representation perturbation and homogeneous output-only distillation. The results show that, under the tested budgets, FSW retains statistically detectable watermark signals across multiple challenging configurations, demonstrating its robustness to common model modifications while characterizing its practical robustness boundary.

1 INTRODUCTION

Large language models (LLMs) and multimodal LLMs have advanced rapidly in recent years, driven in part by the widespread release of capable open weight models such as Llama (Meta AI, 2025), Mixtral (Mistral AI, 2025), Qwen (Yang et al., 2025), and Gemma (Team et al., 2025). As these models become easier to download, fine-tune, and redistribute, questions of model ownership and provenance become increasingly important. In practice, model developers face realistic threats including checkpoint leakage, unauthorized redistribution, and model extraction through supervised querying and knowledge distillation, all of which can transfer capabilities while obscuring attribution. Existing provenance and ownership protections for LLMs mainly fall into two categories. The first is content watermarking, which embeds statistically detectable signals into generated texts for provenance tracing without requiring access to model parameters (Kirchenbauer et al., 2023; Dathathri et al., 2024; Zhang et al., 2024). The second is model watermarking, which aims to embed ownership information into model weights or internal representations so that a suspect model can be verified in an ownership dispute (Li et al., 2023; Sander et al., 2024; Adi et al., 2018). While these directions are complementary, both face important limitations in adversarial settings. Content watermarking is primarily designed to attribute outputs and is vulnerable to strong extraction pipelines, especially when an adversary distills a student model that reproduces capabilities without preserving the original output-level signature, as illustrated in Figure 1. By contrast, robust model-

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.

side watermarking for LLMs remains relatively underexplored, particularly under post-hoc model modifications.

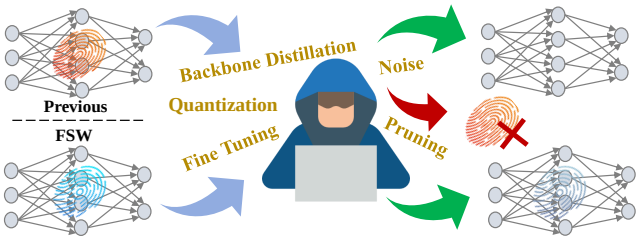


Figure 1: Conceptual comparison of watermark robustness under common model-side attacks. Conventional internal watermarks are often erased by post-hoc modifications, whereas FSW preserves detectability by anchoring the watermark in a stable functional backbone.

To address this gap, we propose **FSW**, a model-side watermarking framework that embeds ownership signals into a model's functional backbone. FSW identifies a low-dimensional subspace that is both task-critical and stable under compression-related transformations, and uses a generalized eigenvalue formulation to separate functional sensitivity from compression-induced variance, yielding stable watermark-bearing directions. We further introduce adaptive spectral truncation to select a spectral band that balances robustness and utility. Experiments show that FSW maintains high watermark verifiability with competitive utility under common function-preserving modifications, including fine-tuning, quantization, pruning, and backbone-preserving distillation. We also evaluate untargeted representation perturbation and homogeneous output-only distillation as stronger stress tests to characterize the empirical robustness boundary of **FSW**.

Our contributions are threefold. First, we propose **FSW**, a model-side watermarking framework that embeds ownership signals into a functional backbone of LLM representations that is both task-critical and modification-stable. By formulating functional-subspace selection as a generalized eigenvalue problem, FSW identifies watermark directions that better balance robustness and utility under compression-related transformations. Second, we develop an orthogonal-key-based multi-bit verification scheme that casts ownership verification as a statistical hypothesis test, enabling reliable watermark recovery while maintaining language-modeling and downstream-task utility. Third, we present extensive experiments across multiple LLM backbones and datasets, showing stable watermark detectability under LoRA fine-tuning, noise perturbation, pruning, quantization, and backbone-preserving distillation. We further examine untargeted representation perturbation and homogeneous output-only distillation to assess how the watermark behaves under stronger modifications and to characterize the practical robustness boundary of the proposed method.

2 Related Work

Content Watermarking. Content watermarking embeds statistically detectable signals into generated texts for provenance tracing without model-parameter access. Early and deployment-oriented methods mainly operate at generation time, including greenlist-based token selection (Kirchenbauer et al., 2023), scalable watermarking systems (Dathathri et al., 2024), and extensions improving robustness, payload, and sampling efficiency (Zhang et al., 2024; Mao et al., 2025; Niess & Kern, 2025). More recent work strengthens verification under post-generation edits through semantic carriers (Dabiriaghdam & Wang, 2025), LLM-paraphraser-based multi-bit recovery (Xu et al., 2025), ensemble-style signals (Niess & Kern, 2025), in-context watermarking (Liu et al., 2025), and distribution-adaptive guarantees (He et al., 2024). These methods are complementary to FSW, but they primarily attribute generated outputs and may also affect truthfulness, safety, and overrefusal (Verma et al., 2025).

Model Watermarking. Model watermarking embeds ownership signals into model parameters or internal representations for verifying suspect models after post-hoc modification. Classical white-box approaches include trigger-set/backdoor ownership proofs (Adi et al., 2018), end-to-end watermark embedding (Darvish Rouhani et al., 2019), and covert robust watermarking for deep networks (Wang & Kerschbaum, 2021). For LLMs, model-side watermarking has been studied through weight-quantization signals (Li et al., 2023), radioactive training traces (Sander et al., 2024), latent-

subspace anchoring such as SEAL (Dai et al., 2025), and black-box ownership verification under model merging (Yamabe et al., 2025). However, robustness under strong extraction, especially knowledge distillation, remains insufficiently understood. Unlike SEAL-style latent anchoring, FSW selects watermark directions via a Fisher-sensitivity/compression-invariance GEVP, explicitly targeting modification-stable functional subspaces for distillation-relevant robustness.

3 Threat Model

We consider a model ownership protection setting in which a model owner releases a watermarked LLM and later encounters a suspect model of disputed provenance. The adversary aims to remove or invalidate the embedded signature while preserving practical utility, as measured by perplexity and downstream task accuracy (Wang & Kerschbaum, 2021; Xu et al., 2024).

Adversary Capabilities and Modification Scope. Following prior model-lineage auditing settings (Shao et al., 2025), we assume a strong white-box adversary with access to the released model's parameters and intermediate representations, but not to the owner's private verification key (Pasquini et al., 2025). The adversary may apply post-hoc training or compression while seeking to preserve task-level utility (Xu et al., 2024; Yan et al., 2023), including supervised fine-tuning, LoRA adaptation, noise perturbation, pruning, quantization, and backbone-preserving distillation. Beyond these common modifications, we evaluate public-information subspace reconstruction, untargeted representation perturbation, and homogeneous output-only distillation as stronger empirical stress tests. In the last setting, a clean-base student with the same architecture and size is trained solely from the watermarked teacher's outputs, without inheriting its parameters or internal representations.

Verification Access and Secret Components. Verification assumes that the owner has white-box access to the suspect model (Wang & Kerschbaum, 2021; Li et al., 2025; Yan et al., 2023). The FSW algorithm and hyperparameters may be public, while the key vectors $\{b_j\}_{j=1}^{M}$, their target-bit mapping $\{y_j\}_{j=1}^{M}$, and the challenge set $\mathcal{C}$ remain secret. The functional basis $U^\star$ and reference mean $\mu_{\mathcal{D}}$ are verifier-side metadata rather than cryptographic secrets: knowing the subspace alone does not reveal the information required for verification. Appendix F evaluates reconstruction and suppression of this subspace using public algorithms and independent data.

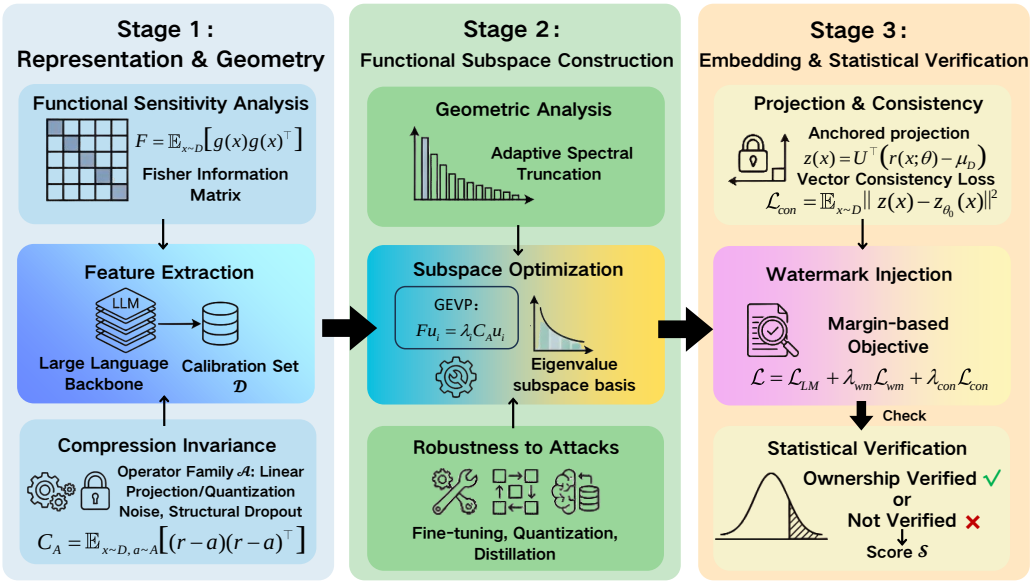


Figure 2: Overall framework of FSW.

4 METHODOLOGY

We propose **FSW**, a white-box watermarking framework that embeds ownership signatures into a stable, low-dimensional functional subspace of LLM representations. To preserve detectability under post-hoc model modifications, FSW identifies directions that are both task-relevant and modification-stable, and embeds and verifies watermarks in the resulting anchored coordinates. As illustrated in Figure 2, FSW comprises three stages: representation geometry estimation, functional subspace construction, and watermark embedding with statistical verification.

4.1 PROBLEM SETUP

Let f_{θ_0} denote a frozen reference autoregressive language model with L transformer layers and hidden dimension d, and let f_θ denote the trainable model initialized from f_{θ_0} for watermark embedding. For an input sequence $x = (x_1, \ldots, x_T)$, we select a fixed internal layer $\ell \in \{1, \ldots, L\}$ and denote its last-token hidden representation by

$$r(x; \theta) = H_T^{(\ell)} \in \mathbb{R}^d. \tag{1}$$

We use $\mathcal{D}$ for the calibration set used in subspace estimation and $\mathcal{C}$ for the challenge set used in watermark embedding and verification. We use k for the functional subspace dimension and M for the number of embedded watermark bits. We use the last-token representation at an intermediate layer as the default watermark interface, since it provides a compact semantic bottleneck for next-token prediction while remaining sufficiently stable for watermarking. Additional notation and representation details are provided in Appendix A.1.

4.2 FUNCTIONAL SUBSPACE CONSTRUCTION

FSW constructs the functional subspace by jointly modeling *functional sensitivity* and *compression invariance*. We estimate a Fisher-style sensitivity matrix on the frozen reference model using $\mathcal{D}$:

$$F = \mathbb{E}_{x \sim \mathcal{D}}[g(x)g(x)^\top], \qquad g(x) = \nabla_r L_{\mathrm{LM}}(x; \theta_0), \tag{2}$$

where the gradient is evaluated with respect to the reference representation $r(x; \theta_0)$. Thus, $u^\top F u$ measures how strongly perturbations along direction u affect the language modeling objective.

To capture modification stability, we model a family of compression-like operators $\mathcal{A}$ and define the corresponding compression-invariance matrix:

$$C_{\mathcal{A}} = \mathbb{E}_{x \sim \mathcal{D},\, a \sim \mathcal{A}}\Big[\big(r(x; \theta_0) - a(r(x; \theta_0))\big)\big(r(x; \theta_0) - a(r(x; \theta_0))\big)^\top\Big]. \tag{3}$$

Here, $u^\top C_{\mathcal{A}} u$ measures the expected distortion of direction u under the modeled operator family. The specific operator instantiations used to define $\mathcal{A}$ are described in Appendix A.2.

We use the following generalized Rayleigh quotient to order candidate directions:

$$\frac{u^\top F u}{u^\top C_{\mathcal{A}} u}. \tag{4}$$

The stationary directions of this quotient satisfy the generalized eigenvalue problem

$$F u_i = \lambda_i C_{\mathcal{A}} u_i. \tag{5}$$

The generalized eigenvalue λ_i can be interpreted as a signal-to-distortion ratio: larger values indicate greater functional sensitivity relative to modeled modification-induced distortion. Numerical details of solving the GEVP are provided in Appendix A.3.

Rather than directly selecting the top-k directions from this surrogate ordering, we apply adaptive spectral truncation as a validation-guided, utility-aware selection rule and choose a spectral band:

$$\mathcal{I} = \{i \mid \tau_{\mathrm{lower}} \lambda_1 \le \lambda_i \le \tau_{\mathrm{upper}} \lambda_1\}. \tag{6}$$

We then form the final functional backbone subspace $U^\star \in \mathbb{R}^{d \times k}$ from the top-k admissible directions. This avoids both highly utility-sensitive extreme directions and weak tail directions; additional discussion of adaptive spectral truncation is provided in Appendix A.4. Theoretical interpretation and proofs are deferred to Appendix B.

4.3 Watermark Embedding and Statistical Verification

Given the functional backbone $U^\star$, we define anchored coordinates by

$$z(x) = U^{\star\top}\left(r(x;\theta) - \mu_{\mathcal{D}}\right) \in \mathbb{R}^k, \tag{7}$$

where $\mu_{\mathcal{D}}$ is the global mean representation of the frozen reference model on $\mathcal{D}$. We correspondingly define the reference coordinates as

$$z_{\theta_0}(x) = U^{\star\top}\left(r(x;\theta_0) - \mu_{\mathcal{D}}\right). \tag{8}$$

To preserve the original model behavior, we impose a representation-level consistency term:

$$L_{\mathrm{con}} = \mathbb{E}_{x\sim\mathcal{D}}\left\|z(x) - z_{\theta_0}(x)\right\|_2^2. \tag{9}$$

For watermark embedding, the model owner samples M approximately orthogonal secret key vectors $\{b_j\}_{j=1}^M \subset \mathbb{R}^k$ and encodes the watermark as target signs $y_j \in \{-1, 1\}$. We use the margin-based objective:

$$L_{\mathrm{wm}} = \mathbb{E}_{x\sim\mathcal{C}} \sum_{j=1}^{M} \max\left(0, \gamma - y_j \frac{b_j^\top z(x)}{\|b_j\|_2}\right), \tag{10}$$

and optimize the final model with

$$L = L_{\mathrm{LM}} + \lambda_{\mathrm{wm}} L_{\mathrm{wm}} + \lambda_{\mathrm{con}} L_{\mathrm{con}}, \tag{11}$$

where L_{LM} is the standard next-token language modeling objective on the training data, λ_{wm} controls watermark strength, and λ_{con} controls representation-level consistency.

At verification time, the defender computes the aggregated watermark score:

$$S = \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} y_j \frac{b_j^\top z(x)}{\|b_j\|_2}. \tag{12}$$

Under the null hypothesis that the suspect model does not contain the target watermark, we approximate

$$S \sim \mathcal{N}(0, \sigma_0^2), \tag{13}$$

where σ_0 is estimated empirically from null-calibration trials such as clean references or mismatched keys. These calibration trials are kept separate from the held-out null controls used to evaluate empirical false positives. Under this Gaussian approximation, the nominal false-positive probability for threshold T is

$$\mathrm{FPR}_{\mathrm{nom}} = \frac{1}{2}\mathrm{erfc}\left(\frac{T}{\sqrt{2}\sigma_0}\right), \tag{14}$$

where $\mathrm{erfc}(\cdot)$ denotes the complementary error function. It is important to distinguish the nominal significance level α derived from the Gaussian tail model from the empirical false positive rate. We empirically validate this threshold selection on 42 independent null trials—including clean models, mismatched keys, and unwatermarked models subjected to modifications—achieving an empirical FPR of 0/42 (detailed in Appendix E).

For multi-bit recovery, we compute per-bit statistics:

$$S_j = \frac{1}{|\mathcal{C}|} \sum_{x\in\mathcal{C}} \frac{b_j^\top z(x)}{\|b_j\|_2}, \qquad \hat{y}_j = \mathrm{sign}(S_j). \tag{15}$$

Extended verification details, ECC usage, and the full procedure are deferred to Appendix C.

5 Experiments

5.1 Experimental Setting

Models and Datasets. We evaluate **FSW** on LLaMA-2-7B-hf (Touvron et al., 2023), Meta-LLaMA-3-8B (Roque, 2024), Qwen2.5-7B (Team et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and

DeepSeek-LLM-7B-Chat (Bi et al., 2024). WikiText-2 (Merity et al., 2016) is used on disjoint splits for estimating F and $C_{\mathcal{A}}$, watermark embedding, and perplexity evaluation; C4 (Raffel et al., 2020) is used for corpus-shift and representation-space stress tests. HellaSwag (Zellers et al., 2019) and ARC-Easy (Clark et al., 2018) evaluate downstream utility. Full dataset configurations are provided in Appendix D.

Baselines and Distillation. We compare FSW with Clean FT, EmMark (Zhang & Koushanfar, 2024), Weighted Quantization (Li et al., 2023), SEAL (Dai et al., 2025), and Naive Top-k, which selects the k largest generalized-eigenvalue directions without adaptive truncation. FSW and SEAL use their intended 8-bit and 64-bit payloads; because their watermark constructions and decision rules differ, we compare bit accuracy, model utility, and native verification outcomes rather than raw detection scores. We evaluate both backbone-preserving distillation, which maintains backbone alignment, and homogeneous output-only distillation, which trains a same-architecture clean-base student solely from teacher outputs.

Metrics and Reporting. Model utility is measured by perplexity, HellaSwag, and ARC-Easy, while watermark reliability is measured by detection score S, bit accuracy, AUC, and the decision $S > T_\alpha$. Unless otherwise specified, verification uses the nominal level $\alpha = 10^{-6}$. Experiments repeated over three runs report mean $\pm$ standard deviation, and cross-method decisions follow each method's native verification rule.

Implementation Details. We use the last-token representation from an intermediate layer. The invariance matrix $C_{\mathcal{A}}$ is estimated using Linear Projection, Quantization Noise, and Structural Dropout, while k and the spectral thresholds are selected under validation-set perplexity constraints. Full training configurations and computational costs are reported in Appendix D.3 and Appendix H, Table 22.

5.2 Main Results

Table 1: Main results across diverse LLM backbones. FSW achieves strong detectability while maintaining competitive functional utility on most evaluated backbones.

Model	Method	Functional Preservation				Detectability		
		PPL↓	ΔPPL↓	HellaSwag↑	ARC-E↑	Det. Score↑	Bit Acc↑	AUC↑
LLaMA-2-7B-hf	Clean FT	5.30	0.00	66.40	71.60	0.47	–	–
	FSW (Ours)	5.91	+0.61	64.80	71.00	**6.09**	**100%**	0.894
LLaMA-3-8B	Clean FT	5.83	0.00	68.60	79.60	-0.03	–	–
	FSW (Ours)	6.20	+0.37	67.80	75.80	**4.00**	**100%**	0.899
Qwen2.5-7B	Clean FT	6.35	0.00	67.00	75.80	1.37	–	–
	FSW (Ours)	7.24	+0.89	66.80	73.40	**5.09**	**100%**	0.977
Mistral-7B-v0.3	Clean FT	16.68	0.00	68.60	77.00	-0.22	–	–
	FSW (Ours)	17.25	+0.57	69.00	77.00	**3.84**	**100%**	0.999
DeepSeek-7B-Chat	Clean FT	59.52	0.00	67.40	72.20	-0.09	–	–
	FSW (Ours)	77.31	+17.79	60.40	66.20	**3.75**	**100%**	1.000

Cross-Backbone Performance We begin by asking whether FSW can preserve ownership detectability across diverse LLM backbones without incurring prohibitive utility loss. Table 1 gives a positive answer: across all five evaluated backbones, FSW achieves 100% bit recovery accuracy with positive detection scores and strong AUC values. Utility degradation is moderate on LLaMA-2-7B-hf, LLaMA-3-8B, Qwen2.5-7B, and Mistral-7B-v0.3, while DeepSeek-7B-Chat exhibits a more pronounced utility–robustness trade-off, suggesting architecture-dependent watermark sensitivity.

Appendix G, Figure 5 further visualizes this cross-backbone behavior. On most models, FSW keeps downstream performance close to the clean baselines while maintaining positive detection scores, indicating that the functional subspace provides verifiable watermark-bearing directions without substantially disturbing task-relevant behavior. Even on DeepSeek-7B-Chat, where utility degradation is larger, the watermark remains recoverable with perfect bit accuracy.

The supplementary tables provide additional support. The cross-corpus evaluation in Appendix H, Table 16 shows that replacing the embedding fine-tuning corpus from WikiText-2 with C4 still preserves 100% bit recovery and near-perfect AUC. The significance analysis in Table 17 further shows that the detection score remains above the critical threshold across significance levels, confirming that the results are not driven by a single corpus or threshold choice.

Robustness under Common Model-Side Modifications We next examine whether FSW remains reliable under common post-hoc modifications that approximately preserve the model's functional backbone. Table 2 gives a compact summary across five LLM backbones. Across LoRA fine-tuning, noise perturbation, pruning, quantization, and backbone-preserving distillation, FSW retains substantial watermark-related scores and bit recovery: the average score retention ranges from 73.04% to 102.99%, while the minimum bit accuracy ranges from 62.50% to 100.00%. A clear pattern also emerges across attack types. Noise preserves 100% bit accuracy with negligible perplexity change, while pruning and quantization introduce larger perplexity changes but retain high average score retention. By average score retention and minimum bit accuracy, backbone-preserving distillation is the most challenging in-scope modification; nevertheless, positive post-modification scores are preserved across all five backbones in the full per-model results reported in Appendix H, Table 12. This is consistent with the design goal of FSW, which targets robustness under modifications that largely preserve the underlying functional backbone.

Table 2: Compact robustness summary under model-side attacks. Results are aggregated over five LLM backbones: LLaMA-2-7B-hf, Meta-LLaMA-3-8B, Qwen2.5-7B, Mistral-7B-v0.3, and DeepSeek-LLM-7B-Chat. Full per-backbone results are provided in Appendix H, Table 12.

Attack	**Watermark Reliability**			**Model Quality**	
	Avg. Ret.↑	**Min Ret.↑**	**Min Bit Acc.↑**	**Avg. ΔPPL↓**	**Max ΔPPL↓**
Backbone Distillation	73.04%	44.17%	62.50%	-5.25	+0.42
LoRA FT	80.16%	17.60%	87.50%	-3.25	-0.10
Noise	97.92%	92.02%	100.00%	-0.01	+0.02
Pruning	93.38%	82.03%	87.50%	+1.04	+4.48
Quantization	102.99%	98.78%	87.50%	+3.83	+15.45

Direct Comparison with SEAL We further compare FSW with SEAL (Dai et al., 2025) using aligned data, utility evaluation, and attack settings, while retaining each method's intended payload and native verification rule. Both methods remain verifiable under LoRA fine-tuning, quantization, noise, and pruning, with the complete results reported in Appendix H, Table 13. The clearest difference appears under backbone-preserving distillation: FSW retains 87.5% and 75.0% bit accuracy on LLaMA-2 and Qwen2.5, respectively, and remains detectable at the mean level, whereas SEAL drops to 53.1% and 51.6% and passes its native verification in 0/3 runs on both backbones. This difference is not achieved through a higher no-attack perplexity. Under Hom.-500, FSW remains detectable on LLaMA-2 but not Qwen2.5, while SEAL passes 0/3 runs on both backbones, further characterizing clean-student extraction as a robustness boundary.

Table 3: Direct comparison between FSW and SEAL under representative robustness settings. Results are mean ± standard deviation over three runs. FSW and SEAL use their intended 8-bit and 64-bit payloads, respectively. Mean Det. denotes the FSW decision based on the mean detection score, while Native Passes reports the number of SEAL runs passing its native verification rule.Complete results are provided in Appendix H, Table 13.

Model	**Attack**	**FSW (8-bit)**		**SEAL (64-bit)**	
		PPL / Bit Acc.	**Mean Det.**	**PPL / Bit Acc.**	**Native Passes**
LLaMA-2-7B	No Attack	5.961 ± 0.071 / 1.000 ± 0.000	Yes	14.378 ± 0.261 / 1.000 ± 0.000	3/3
	BPD	5.982 ± 0.063 / 0.875 ± 0.125	Yes	9.825 ± 0.112 / 0.531 ± 0.072	0/3
	Hom.-500	7.995 ± 0.338 / 0.625 ± 0.000	Yes	11.524 ± 0.186 / 0.557 ± 0.048	0/3
Qwen2.5-7B	No Attack	7.201 ± 0.132 / 1.000 ± 0.000	Yes	13.408 ± 0.055 / 1.000 ± 0.000	3/3
	BPD	7.664 ± 0.110 / 0.750 ± 0.000	Yes	11.984 ± 0.036 / 0.516 ± 0.031	0/3
	Hom.-500	8.145 ± 0.025 / 0.500 ± 0.000	No	11.884 ± 0.028 / 0.542 ± 0.024	0/3

Robustness Boundary Analysis We further evaluate two stronger stress-test settings. Under untargeted representation perturbations, the detection score varies non-monotonically but remains above the corresponding nominal threshold across all tested budgets. At stronger perturbation levels, bit recovery accuracy generally decreases even when the aggregate detection score increases, while perplexity changes become more visible, indicating that ownership detection and exact payload recovery capture complementary aspects of watermark reliability. Because the attack does not have access to the private key vectors, target-bit mapping, or challenge set, it cannot directly optimize the ownership verification statistic. The complete multi-seed curves and per-seed results are provided in Appendix G, Figure 7, Appendix G, Tables 21 and 23.

We next consider homogeneous unconstrained distillation as a stronger stress test. Table 4 shows that output-only homogeneous distillation substantially weakens the watermark signal across backbones, while still allowing detectable signal to persist under the tested budgets. On LLaMA-2-7B, the detection score drops from 6.094 for the teacher to 2.542 after 500 steps while remaining positive; on Qwen-2.5-7B, the signal is more fragile at early budgets but becomes positive again at longer budgets. The supplementary statistical test in Appendix H, Table 10 further shows that the 500-step homogeneous student remains above the corresponding threshold across all tested significance levels in expectation, with a positive margin even at $\alpha = 10^{-8}$. Taken together, these results show that FSW can still preserve statistically detectable ownership signal under stronger homogeneous distillation, while also characterizing the practical robustness boundary of representation-level watermarking. Appendix H, Table 24 further provides a single-seed diagnostic sweep, and Figure 4 visualizes the same trajectory, showing that student utility recovery and residual watermark strength remain coupled during homogeneous unconstrained distillation.

Table 4: Mean-level verification under homogeneous output-only distillation at the nominal significance level $\alpha = 10^{-6}$. Values are mean $\pm$ std. over three runs; detection is determined by the sign of the mean margin $S - T_\alpha$. Complete results are provided in Appendix H, Table 14.

Backbone	**Steps**	**PPL**$\downarrow$	**Score** $S\uparrow$	$T_{10^{-6}}$	**Margin** $S - T$	**Detected**
LLaMA-2-7B	500	7.995 ± 0.338	2.542 ± 1.010	1.954	0.588 ± 1.010	Yes
	1000	7.098 ± 0.123	2.417 ± 0.048	1.954	0.463 ± 0.048	Yes
	1500	6.767 ± 0.034	1.706 ± 0.711	1.954	-0.248 ± 0.711	No
Qwen2.5-7B	500	8.145 ± 0.025	-1.813 ± 0.299	1.024	-2.837 ± 0.299	No
	1000	7.848 ± 0.012	1.040 ± 0.412	1.024	0.016 ± 0.412	Yes
	1500	7.838 ± 0.017	1.656 ± 0.521	1.024	0.632 ± 0.521	Yes

5.3 Ablation and Sensitivity Analysis

Effect of Spectral Direction Selection We next examine whether the way carrier directions are selected from the generalized spectrum is critical to robustness. Appendix H, Table 11 gives a clear answer. Compared with fixed alternatives such as Top-k, Mid-50%, Low-75%, and Last-k, adaptive spectral truncation achieves the strongest overall post-attack performance under backbone distillation. In particular, it yields the highest post-attack detection score, the smallest score drop, and the best score retention, while also maintaining the lowest post-attack perplexity among all compared choices. These results show that robust watermark carriers should not be selected simply from the largest generalized eigenvalues or from arbitrarily chosen spectral regions.

Figure 3 further clarifies this pattern. The left panel shows that adaptive selection preserves the highest post-attack score while starting from the strongest pre-attack baseline, and the right panel shows that it also achieves the highest retention among all alternatives. By contrast, Top-k and tail-biased selections either sacrifice post-attack robustness or fail to maintain equally strong pre-attack detectability. This is consistent with the design principle of FSW: the most robust watermark directions lie in a balanced spectral band that jointly preserves functional sensitivity and modification stability, rather than at either extreme of the generalized spectrum. Taken together, these results justify adaptive spectral truncation as a necessary design choice for constructing a stable watermark-bearing functional subspace.

Capacity and Subspace Size. We provide detailed payload, embedding-dimension, and ECC analyses in Appendix G and Appendix H. Figure 6a shows that moderate payloads provide the best utility–detectability trade-off, while overly large payloads can degrade model utility even when detection

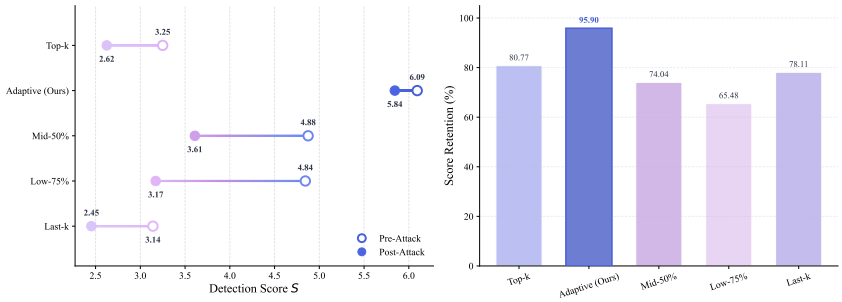


Figure 3: Spectral direction selection ablation under backbone distillation. Adaptive spectral truncation preserves the strongest post-attack score retention while keeping post-attack perplexity low.

remains strong. Figure 6(b), together with Table 19 and Table 20, shows that the best embedding dimension k is architecture-dependent: too small a subspace may miss stable watermark-bearing directions, whereas too large a subspace may include fragile or utility-sensitive directions. Table 15 further shows that lightweight ECC can correct residual bit errors in moderate-capacity regimes.

Table 5: Component ablation on Llama-2-7B-hf. *Pre* and *Post* denote metrics before and after the robustness attack; each variant removes one FSW component.

Variant	**Detection Score S ↑**			**Bit Accuracy ↑**			**Utility (PPL) ↓**		
	Pre	Post	ΔS	Pre	Post	ΔAcc	Pre	Post	ΔPPL
Full FSW	**6.09**	**5.84**	**-0.25**	**100.0%**	**87.5%**	**-12.5%**	**5.91**	**5.91**	**0.00**
w/o Comp. Inv.	4.47	0.96	-3.51	100.0%	100.0%	0.00%	7.02	6.43	+0.60
w/o Anchored Proj.	3.08	-1.94	-5.02	75.0%	37.5%	-37.5%	7.31	6.82	+0.49
w/o Consistency Loss	4.59	-0.36	-4.95	100.0%	37.5%	-62.5%	7.19	6.59	+0.60
w/o Adaptive Spec. Trunc.	4.91	-2.73	-7.64	87.5%	12.5%	-75.0%	7.16	6.85	+0.30

Component Ablation We next examine whether the key design components of FSW are all necessary for robustness and utility preservation. Table 5 gives a clear answer. The full FSW configuration achieves the best overall trade-off, maintaining a high post-attack detection score of 5.84 with stable perplexity at 5.91. Removing any single component degrades robustness, showing that the effectiveness of FSW does not come from any one mechanism alone. A clear pattern also emerges across variants. Removing compression invariance weakens post-attack detectability and increases perplexity, while removing anchored projection or consistency loss causes substantially larger drops in post-attack detection. Most notably, removing adaptive spectral truncation leads to the most severe degradation, driving the post-attack detection score to -2.73 and reducing post-attack bit accuracy to 12.5%. Taken together, these results show that the robustness of FSW relies on the joint effect of compression invariance modeling, anchored projection, consistency preservation, and adaptive spectral truncation.

6 Conclusion

We propose **FSW**, a model-side watermarking framework for LLM ownership verification. **FSW** embeds watermark signals into a stable functional subspace, preserving detectability under common functionality-preserving model modifications while maintaining competitive functional utility across most evaluated backbones. Stronger stress-test settings, including public-information subspace reconstruction, untargeted representation perturbation, and homogeneous output-only distillation, characterize the empirical robustness boundary of representation-level watermarking. **FSW** provides a principled and practical foundation for white-box ownership verification, with future extensions toward black-box verification, heterogeneous distillation, and architecture-mismatched extraction. Future work will further evaluate its impact on instruction following, long-context behavior, open-ended generation quality, and safety alignment.

AI USE STATEMENT

Generative AI tools were used solely to aid and polish author-written text by improving grammar, clarity, and concision. They were not used for literature retrieval, research ideation or execution, drafting scientific content, synthetic-data generation, mathematical proofs, or the generation and interpretation of experimental results. All AI-assisted edits were reviewed and approved by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies watermarking for model provenance and ownership verification. The experiments use publicly available pretrained models and standard research datasets, and involve no human participants or newly collected personal data. The proposed method is intended for defensive provenance auditing, and the attack evaluations are included to characterize its empirical security boundaries. Watermark evidence should not be treated as the sole legal proof of model ownership, and deployment should comply with applicable licenses, policies, and laws.

REPRODUCIBILITY STATEMENT

The complete formulation and verification procedure of FSW are provided in Section 4, and the evaluation protocol is described in Section 5.1. Dataset splits, implementation and hardware configurations, hyperparameters, attack and distillation settings, and complete experimental results are provided in Appendices D, D.3, and H.

REFERENCES

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pp. 1615–1631, 2018.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

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

Amirhossein Dabiriaghdam and Lele Wang. SimMark: A robust sentence-level similarity-based watermarking algorithm for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30785–30806, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1567. URL https://aclanthology.org/2025.emnlp-main.1567/.

Yanbo Dai, Zongjie Li, Zhenlan Ji, and Shuai Wang. Seal: Subspace-anchored watermarks for llm ownership. *arXiv preprint arXiv:2511.11356*, 2025.

Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems*, pp. 485–497, 2019.

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

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Theoretically grounded framework for llm watermarking: A distribution-adaptive approach. In *The 1st Workshop on GenAI Watermarking*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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

Linyang Li, Botian Jiang, Pengyu Wang, Ke Ren, Hang Yan, and Xipeng Qiu. Watermarking llms with weight quantization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3368–3378, 2023.

Zongjie Li, Daoyuan Wu, Shuai Wang, and Zhendong Su. Differentiation-based extraction of proprietary data from fine-tuned llms. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3071–3085, 2025.

Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. In-context watermarks for large language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Minjia Mao, Dongjun Wei, Zeyu Chen, Xiao Fang, and Michael Chau. Watermarking large language models: An unbiased and low-risk method. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7939–7960, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

Mistral AI. Introducing mistral 3. https://mistral.ai/news/mistral-3, December 2025.

Georg Niess and Roman Kern. Ensemble watermarks for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2903–2916, 2025.

Dario Pasquini, Evgenios M Kornaropoulos, and Giuseppe Ateniese. {LLMmap}: Fingerprinting for large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 299–318, 2025.

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

Luís Roque. The evolution of llama: From llama 1 to llama 3.1. *Medium*, 2024.

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

Shuo Shao, Yiming Li, Yu He, Hongwei Yao, Wenyuan Yang, Dacheng Tao, and Zhan Qin. Sok: Large language model copyright auditing via fingerprinting. *arXiv preprint arXiv:2508.19843*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry,

Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

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

Apurv Verma, NhatHai Phan, and Shubhendu Trivedi. Watermarking degrades alignment in language models: Analysis and mitigation. *arXiv preprint arXiv:2506.04462*, 2025.

Tianhao Wang and Florian Kerschbaum. Riga: Covert and robust white-box watermarking of deep neural networks. In *Proceedings of the web conference 2021*, pp. 993–1004, 2021.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024.

Xiaojun Xu, Jinghan Jia, Yuanshun Yao, Yang Liu, and Hang Li. Robust multi-bit text watermark with llm-based paraphrasers. In *Forty-second International Conference on Machine Learning*, 2025.

Shojiro Yamabe, Futa Kai Waseda, Tsubasa Takahashi, and Koki Wataoka. Mergeprint: Merge-resistant fingerprints for robust black-box ownership verification of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6894–6916, 2025.

Yifan Yan, Xudong Pan, Mi Zhang, and Min Yang. Rethinking {White-Box} watermarks on deep learning models under neural structural obfuscation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2347–2364, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Ruisi Zhang and Farinaz Koushanfar. Emmark: Robust watermarks for ip protection of embedded quantized large language models. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–6, 2024.

Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, and Farinaz Koushanfar. {REMARK-LLM}: A robust and efficient watermarking framework for generative large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1813–1830, 2024.

A Notation and Additional Implementation Details

Notation Summary. For ease of reference, Table 6 summarizes the main symbols used in FSW.

Table 6: Summary of the main symbols used in FSW.

Symbol	Meaning
L	Total number of transformer layers
ℓ	Selected watermark layer index
d	Hidden dimension of the selected layer
$x = (x_1, \ldots, x_T)$	Input sequence
$r(x; \theta)$	Last-token hidden representation at layer ℓ
$\mathcal{D}$	Calibration set for subspace estimation
$\mathcal{C}$	Challenge set for watermark embedding and verification
$\mathcal{A}$	Family of compression-like operators
F	Fisher-style functional sensitivity matrix
$C_{\mathcal{A}}$	Compression-invariance matrix under operator family $\mathcal{A}$
$U^\star \in \mathbb{R}^{d \times k}$	Functional backbone subspace basis
k	Functional subspace dimension
$z(x)$	Anchored projected representation in the functional subspace
$\mu_{\mathcal{D}}$	Mean representation of the frozen reference model on $\mathcal{D}$
M	Number of embedded watermark bits
b_j	Secret key vector for the j-th watermark bit
$y_j \in \{-1, 1\}$	Target sign of the j-th watermark bit
γ	Margin in the watermark embedding objective
λ_{wm}	Weight of the watermark embedding loss
λ_{con}	Weight of the consistency loss
S	Aggregated detection score
σ_0	Null standard deviation in hypothesis testing
T_α	Detection threshold at significance level α

A.1 Choice of Representation Interface

We use the last-token representation because, in autoregressive language models, it provides a direct semantic bottleneck for next-token prediction while avoiding ambiguity from token aggregation. This yields a compact single-vector interface for watermark embedding and verification.

Unless otherwise noted, we use an intermediate layer as the default watermark layer. Earlier layers are often more syntactic and less semantically consolidated, whereas later layers are more tightly coupled to utility-sensitive decoding features. Intermediate layers therefore provide a practical stability-expressiveness trade-off; we leave a systematic layer-wise study to future work.

A.2 Compression Operator Family

To capture stability under common post-hoc model modifications, we instantiate the operator family $\mathcal{A}$ using three representative compression-like perturbations:

- *Linear Projection:* $a(r) = W_{\mathrm{low}} r$, simulating capacity bottlenecks.
- *Quantization Noise:* $a(r) = r + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 I)$, simulating precision loss.
- *Structural Dropout:* $a(r) = r \odot m$, simulating feature sparsification.

This construction targets functionality-preserving modifications and should not be interpreted as claiming robustness to fully unconstrained distillation with arbitrary latent-space reparameterization.

A.3 Numerical Solution of the GEVP

In practice, the generalized eigenvalue problem

$$Fu_i = \lambda_i C_{\mathcal{A}} u_i$$

is solved on a numerically stable subspace of $C_{\mathcal{A}}$. When needed, we use a regularized form

$$C_{\mathcal{A}}^{(\epsilon)} = C_{\mathcal{A}} + \epsilon I$$

with a small $\epsilon > 0$ to avoid instability caused by near-singular directions.

A.4 Adaptive Spectral Truncation

Adaptive spectral truncation selects directions from a bounded spectral band

$$\mathcal{I} = \{i \mid \tau_{\text{lower}} \lambda_1 \leq \lambda_i \leq \tau_{\text{upper}} \lambda_1\}.$$

This design avoids both extremes of the generalized spectrum:

- Very large generalized eigenvalues may correspond to highly utility-sensitive semantic directions.
- Very small generalized eigenvalues often correspond to weak or unstable directions that are easier to erase.

The final functional backbone subspace $U^\star$ is formed by selecting the top-k admissible directions within this spectral band.

B Theoretical Interpretation and Proofs

In this appendix, we provide proof sketches for the theoretical results underlying the functional subspace construction in Section 4. These results clarify why the generalized eigenspace of $(F, C_{\mathcal{A}})$ provides a natural functional subspace for watermark embedding, and how compression-induced instability propagates to the final verification statistic.

B.1 Proof of Proposition 1

Proposition 1 (Optimality under Trace Surrogate)**.** *Assume that F is symmetric and that $C_{\mathcal{A}}$ is positive definite on the numerically retained subspace, or is replaced by its regularized form. For any nonzero direction $u \in \mathbb{R}^d$, define its functional utility as $u^\top F u$ and its modification instability as $u^\top C_{\mathcal{A}} u$. The stationary directions of the generalized Rayleigh quotient*

$$\mathcal{R}(u) = \frac{u^\top F u}{u^\top C_{\mathcal{A}} u}$$

are exactly the generalized eigenvectors satisfying

$$Fu = \lambda C_{\mathcal{A}} u.$$

Let $\lambda_1 \geq \cdots \geq \lambda_d$ denote the generalized eigenvalues and let the corresponding eigenvectors be $C_{\mathcal{A}}$-orthonormal. Then $U_k = [u_1, \ldots, u_k]$ maximizes

$$\operatorname{tr}(U^\top F U) \quad \textit{subject to} \quad U^\top C_{\mathcal{A}} U = I.$$

Thus, the top-k generalized eigenvectors provide an optimal solution to this trace surrogate without additional embedding-utility constraints.

Proof. Because $\mathcal{R}(u)$ is invariant to the scale of u, we impose the normalization constraint

$$u^\top C_{\mathcal{A}} u = 1$$

and maximize $u^\top F u$. Introducing a Lagrange multiplier λ gives

$$\mathcal{L}(u, \lambda) = u^\top F u - \lambda\big(u^\top C_{\mathcal{A}} u - 1\big).$$

Setting the derivative with respect to u to zero yields

$$2Fu - 2\lambda C_{\mathcal{A}} u = 0,$$

and therefore

$$Fu = \lambda C_{\mathcal{A}} u.$$

Conversely, every normalized generalized eigenvector satisfies this stationarity condition.

Since $C_{\mathcal{A}}$ is positive definite on the considered subspace, the generalized eigenvectors can be chosen to be $C_{\mathcal{A}}$-orthonormal. The generalized Courant–Fischer characterization then gives

$$\max_{U^\top C_{\mathcal{A}} U = I} \operatorname{tr}(U^\top F U) = \sum_{i=1}^{k} \lambda_i,$$

which is attained by $U_k = [u_1, \ldots, u_k]$. Hence, the top-k generalized eigenvectors maximize the stated trace surrogate. □

Remark on Adaptive Spectral Truncation. The top-k optimality above applies only to the trace surrogate without the spectral-band and embedding-utility constraints used by FSW. Directions associated with the largest generalized eigenvalues may be highly utility-sensitive, whereas directions near the spectral tail may provide weak or unstable functional signals. Adaptive Spectral Truncation therefore restricts selection to the band $\mathcal{I}$ as an empirically motivated, utility-aware rule rather than a closed-form global optimum of the final embedding objective. Proposition 2 and Corollary 1 depend only on the projected instability of the final selected subspace U and therefore continue to hold for the adaptively selected subspace.

B.2 Proof of Proposition 2

Proposition 2 (Projected instability of a functional subspace)**.** *Let $U \in \mathbb{R}^{d\times k}$ be any orthonormal subspace basis, and let $\Delta r(x,a) = r(x) - a(r(x))$. Then*

$$\mathbb{E}_{x\sim\mathcal{D},\, a\sim\mathcal{A}} \left[\|U^\top \Delta r(x,a)\|_2^2\right] = \operatorname{tr}\big(U^\top C_{\mathcal{A}} U\big).$$

In particular, among subspaces with comparable functional utility, those with smaller $\operatorname{tr}(U^\top C_{\mathcal{A}} U)$ induce smaller expected functional-subspace coordinate drift under the modeled modification family.

Proof. For any fixed (x,a), define $\Delta r = \Delta r(x,a) = r(x) - a(r(x))$. Then

$$\|U^\top \Delta r\|_2^2 = (U^\top \Delta r)^\top (U^\top \Delta r) = \Delta r^\top U U^\top \Delta r.$$

Using the trace identity $v^\top A v = \operatorname{tr}(A v v^\top)$ for any vector v and matrix A, we have

$$\Delta r^\top U U^\top \Delta r = \operatorname{tr}(U U^\top \Delta r \Delta r^\top).$$

Taking expectation over $x \sim \mathcal{D}$ and $a \sim \mathcal{A}$ gives

$$\mathbb{E}_{x,a} \left[\|U^\top \Delta r(x,a)\|_2^2\right] = \operatorname{tr}\big(U U^\top \mathbb{E}_{x,a} \left[\Delta r(x,a) \Delta r(x,a)^\top\right]\big).$$

By the definition of $C_{\mathcal{A}}$,

$$C_{\mathcal{A}} = \mathbb{E}_{x,a} \left[\Delta r(x,a) \Delta r(x,a)^\top\right].$$

Therefore,

$$\mathbb{E}_{x,a} \left[\|U^\top \Delta r(x,a)\|_2^2\right] = \operatorname{tr}(U U^\top C_{\mathcal{A}}).$$

Using cyclic invariance of trace,

$$\operatorname{tr}(U U^\top C_{\mathcal{A}}) = \operatorname{tr}(U^\top C_{\mathcal{A}} U).$$

This proves the identity

$$\mathbb{E}_{x,a} \left[\|U^\top \Delta r(x,a)\|_2^2\right] = \operatorname{tr}(U^\top C_{\mathcal{A}} U).$$

The second claim follows immediately: for subspaces with similar functional relevance, a smaller $\operatorname{tr}(U^\top C_{\mathcal{A}} U)$ means a smaller expected perturbation energy after projection into the functional subspace, and thus a more stable watermark-bearing subspace. □

B.3 Proof of Corollary 1

Corollary 1 (Score drift follows functional-subspace coordinate drift)**.** *Let $z(x)$ and $z'(x)$ denote the anchored functional-subspace coordinates before and after modification. Then the change in the aggregated watermark score satisfies*

$$|S' - S| \le \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} \frac{|b_j^\top (z'(x) - z(x))|}{\|b_j\|_2} \le \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} \|z'(x) - z(x)\|_2.$$

Hence, subspaces with smaller projected instability also tend to yield more stable verification statistics.

Proof. Recall the verification score

$$S = \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} y_j \frac{b_j^\top z(x)}{\|b_j\|_2},$$

and similarly

$$S' = \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} y_j \frac{b_j^\top z'(x)}{\|b_j\|_2}.$$

Subtracting,

$$S' - S = \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} y_j \frac{b_j^\top (z'(x) - z(x))}{\|b_j\|_2}.$$

Taking absolute values and using the triangle inequality,

$$|S' - S| \le \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} \frac{|b_j^\top (z'(x) - z(x))|}{\|b_j\|_2}.$$

Then, by Cauchy–Schwarz,

$$|b_j^\top (z'(x) - z(x))| \le \|b_j\|_2 \, \|z'(x) - z(x)\|_2.$$

Substituting this into the above inequality yields

$$|S' - S| \le \frac{1}{|\mathcal{C}|M} \sum_{x\in\mathcal{C}} \sum_{j=1}^{M} \|z'(x) - z(x)\|_2.$$

This proves the stated bound. Therefore, if the anchored functional-subspace coordinates remain stable under post-hoc modification, then the aggregated verification score also remains stable. In particular, a functional subspace with smaller projected instability yields a smaller upper bound on verification-score drift. □

C Extended Verification Details and Algorithm

C.1 Gaussian Approximation and Threshold Selection

Under the null hypothesis H_0 that the suspect model does not contain the target watermark, we approximate the aggregated score by

$$S \sim \mathcal{N}(0, \sigma_0^2).$$

This Gaussian approximation is motivated by the fact that S aggregates many weakly dependent signed projections across challenge samples and bit directions. In practice, σ_0 is estimated empirically from null trials, such as non-watermarked reference models or randomly mismatched secret keys, rather than assumed from theory alone.

For a threshold T, the false positive rate is

$$\mathrm{FPR} = \frac{1}{2}\mathrm{erfc}\left(\frac{T}{\sqrt{2}\sigma_0}\right).$$

Equivalently, for a target significance level α, we choose a threshold T_α such that

$$\Pr(S > T_\alpha \mid H_0) \le \alpha.$$

C.2 Multi-bit Decoding and ECC

For message recovery, we compute the per-bit statistics

$$S_j = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \frac{b_j^\top z(x)}{\|b_j\|_2}, \qquad \hat{y}_j = \mathrm{sign}(S_j).$$

When stronger robustness is required, the recovered bits can be further decoded using error-correcting codes (ECC). In our experiments, ECC serves as an optional post-processing step and does not affect the definition of the core verification statistic.

C.3 Full Procedure

Algorithm 1 summarizes the complete watermark generation, embedding, and verification pipeline.

D Detailed Experimental Setup

D.1 Model Selection and Specifications

Table 7: Architectural specifications of the evaluated language models. The watermark is consistently embedded in the middle layer ($\ell \approx L/2$) for each architecture to balance representation stability and expressiveness.

Model	**Parameters**	**Hidden Size**	**Layers**	**Watermark Layer** (ℓ)
GPT-2	124M	768	12	8
LLaMA-2-7B-hf	7B	4096	32	16
Meta-LLaMA-3-8B	8B	4096	32	16
Qwen2.5-7B	7B	3584	28	14
Mistral-7B-v0.3	7B	4096	32	16
DeepSeek-7B-Chat	7B	4096	30	15

The model selection is shown in Table 7.

D.2 Dataset Descriptions and Allocation

We combine multiple datasets for subspace estimation, fine-tuning, and functional evaluation to assess whether the watermark remains stable under corpus shift and to avoid coupling subspace construction, watermark embedding, and downstream evaluation.

- **WikiText-2:** A high-quality collection of Wikipedia articles. We use up to 3,000 training sequences for watermark fine-tuning, 500 training samples for estimating each of F and $C_{\mathcal{A}}$, 500 validation samples for PPL/challenge construction, and up to 500 test samples for watermark verification.
- **C4 (Colossal Clean Crawled Corpus):** A massive, cleaned version of Common Crawl. We sample 3,000 sequences for watermark embedding via LoRA fine-tuning to enhance distributional robustness.
- **HellaSwag & ARC-Easy:** Used as downstream benchmarks (500 samples each) to assess common-sense reasoning and knowledge retention after watermarking.

D.3 Implementation Details

The proposed framework is implemented using PyTorch and the PEFT library. We extract the last-token hidden representations from the middle layer of each model as specified in Table 7. For the invariance matrix $C_{\mathcal{A}}$, we apply $N = 3$ stochastic perturbations per sample using a family of operators $\mathcal{A}$:

Algorithm 1 Functional Subspace Watermarking (FSW)

Input: Frozen reference model f_{θ_0}, language-modeling training set $\mathcal{D}_{\mathrm{LM}}$, calibration set $\mathcal{D}$, private challenge set $\mathcal{C}$, operator family $\mathcal{A}$, watermark layer ℓ, subspace dimension k, margin γ, and target bits encoded as signs $\{y_j\}_{j=1}^{M} \subset \{-1, +1\}$.
Hyperparameters: Spectral thresholds $\tau_{\mathrm{lower}}, \tau_{\mathrm{upper}}$, loss weights $\lambda_{\mathrm{wm}}, \lambda_{\mathrm{con}}$, and significance level α.
Step 1: Representation Interface and Geometry Estimation
    Extract last-token hidden states $r(x; \theta_0)$ from layer ℓ for $x \sim \mathcal{D}$ and compute the mean $\mu_{\mathcal{D}}$.
    Estimate F from language-modeling gradients on $\mathcal{D}$, and estimate $C_{\mathcal{A}}$ from the distortions induced by $a \sim \mathcal{A}$ on $\mathcal{D}$.
Step 2: Functional Subspace Construction
    Solve the generalized eigenvalue problem $Fu_i = \lambda_i C_{\mathcal{A}} u_i$.
    Select the admissible spectral band $\mathcal{I}$ using adaptive spectral truncation thresholds $\tau_{\mathrm{lower}}, \tau_{\mathrm{upper}}$.
    Construct the functional backbone basis $U^\star = [u_{i_1}, \ldots, u_{i_k}]$ from the top-k admissible directions.
Step 3: Anchored Watermark Embedding
    Initialize the trainable model parameters as $\theta \leftarrow \theta_0$.
    Generate M approximately orthogonal secret key vectors $\{b_1, \ldots, b_M\} \subset \mathbb{R}^k$.
    Define the anchored coordinates

$$z(x) = U^{\star\top}\big(r(x;\theta) - \mu_{\mathcal{D}}\big), \qquad z_{\theta_0}(x) = U^{\star\top}\big(r(x;\theta_0) - \mu_{\mathcal{D}}\big).$$

    Compute L_{LM} on $\mathcal{D}_{\mathrm{LM}}$, L_{wm} on $\mathcal{C}$, and L_{con} on $\mathcal{D}$.
    Fine-tune f_θ by minimizing

$$L = L_{\mathrm{LM}} + \lambda_{\mathrm{wm}} L_{\mathrm{wm}} + \lambda_{\mathrm{con}} L_{\mathrm{con}}.$$

    Retain $U^\star$, $\mu_{\mathcal{D}}$, $\mathcal{C}$, and $\{(b_j, y_j)\}_{j=1}^{M}$ as verifier-side metadata.
Step 4: Statistical Ownership Verification
    Given a suspect model $f_{\tilde{\theta}}$, compute

$$z_{\tilde{\theta}}(x) = U^{\star\top}\big(r(x;\tilde{\theta}) - \mu_{\mathcal{D}}\big)$$

    for $x \in \mathcal{C}$.
    Estimate σ_0 from separate null-calibration trials and set

$$T_\alpha = \sqrt{2}\sigma_0 \,\mathrm{erfc}^{-1}(2\alpha).$$

    Evaluate the aggregated score S using $z_{\tilde{\theta}}(x)$ and reject H_0 if $S > T_\alpha$.
Step 5: Multi-bit Decoding
    Compute the per-bit statistics S_j using $z_{\tilde{\theta}}(x)$ for $j = 1, \ldots, M$.
    Recover $\hat{y}_j = \mathrm{sign}(S_j)$ and apply ECC decoding if the target signs were obtained from an ECC-encoded message.
    Return: Watermarked model f_θ, verifier metadata, ownership decision $\mathbb{I}[S > T_\alpha]$, recovered target signs $\{\hat{y}_j\}_{j=1}^{M}$, and the optional ECC-decoded message.

- **Linear Projection:** rank ratio of 0.25 relative to the hidden dimension.
- **Quantization Noise:** Gaussian noise with $\sigma = 0.1$.
- **Structural Dropout:** Bernoulli mask with retention rate $p = 0.9$.

The functional backbone $U^\star$ is constructed using the model-specific dimension k selected by the best-k scan (Table 20), using adaptive thresholds $\tau_{\mathrm{lower}} = 10^{-4}$ and $\tau_{\mathrm{upper}} = 0.6$. LoRA fine-tuning uses rank $r = 16$ and $\alpha = 32$. The joint objective is optimized with $\lambda_{\mathrm{wm}} = 10$, $\lambda_{\mathrm{con}} = 0.1$, and a margin $\gamma = 5.0$. All experiments are conducted on a single NVIDIA RTX 6000 Ada GPU.

E Empirical False Positive Rate and Null Calibration

To empirically validate the specificity of FSW and the reliability of the nominal threshold T_α derived from the Gaussian approximation, we conducted 42 independent negative-control (null) trials across LLaMA-2-7B and Qwen2.5-7B. These trials cover clean reference models, models modified by various attacks (BPD, LoRA, Noise, Pruning, Quantization) without the watermark, and trials using mismatched secret keys.

As shown in Table 8, across all 42 null trials, the maximum detection score remained strictly below the corresponding nominal threshold ($T_{10^{-6}} = 1.954$ for LLaMA-2 and 1.024 for Qwen2.5). We observed 0 false positives, yielding an empirical FPR of 0/42 under the tested conditions. This provides strong empirical support for the statistical separability of the functional subspace, confirming that the verification process does not spuriously trigger on unwatermarked or modified models.

Table 8: Empirical False Positive (FP) evaluation across 42 negative-control trials. Values represent the mean $\pm$ standard deviation of the detection score S, with the maximum score observed across 3 runs in parentheses. No false positives were observed.

Null Condition	LLaMA-2 S (max)	FP	Qwen2.5 S (max)	FP
Clean base	-1.807 ± 0.403 (-1.344)	0/3	-0.738 ± 0.463 (-0.332)	0/3
BPD Attack	-0.796 ± 0.417 (-0.336)	0/3	-1.384 ± 0.700 (-0.582)	0/3
LoRA FT	-1.841 ± 0.424 (-1.352)	0/3	-0.714 ± 0.457 (-0.313)	0/3
Noise	-1.893 ± 0.446 (-1.398)	0/3	-0.709 ± 0.359 (-0.338)	0/3
Pruning	0.842 ± 0.448 (1.359)	0/3	-0.349 ± 0.465 (0.044)	0/3
Quantization	-0.239 ± 0.564 (0.404)	0/3	0.530 ± 0.425 (0.832)	0/3
Mismatched key	-2.622 ± 3.776 (1.398)	0/3	-2.712 ± 2.843 (-0.723)	0/3
Overall		**0/21**		**0/21**

F Robustness against Public-Information Subspace-Reconstruction

A strong white-box adversary might attempt to bypass the secret-key verification by reconstructing and suppressing the functional subspace. Assuming the adversary knows the target layer ℓ, subspace dimension k, spectral thresholds, and the public GEVP algorithm, they may use independent public data to re-estimate $\hat{F}$ and $\hat{C}_{\mathcal{A}}$, construct a surrogate subspace $\hat{U}$, and persistently suppress it.

We measure the normalized subspace overlap as $\rho_{\text{overlap}} = \frac{1}{k}\|(U^\star)^\top \hat{U}\|_F^2$, where $\rho_{\text{overlap}} = 1$ denotes identical subspaces. Across three seeds, the measured overlaps are 0.00397 ± 0.00010 for LLaMA-2-7B and 0.00952 ± 0.00031 for Qwen2.5-7B, which are extremely close to their random expectations ($k/d = 0.00391$ and 0.00893, respectively). This indicates that independent public data achieves only random-level alignment with the owner's realized secret subspace.

Furthermore, we evaluated the strongest targeted suppression attack (suppression strength $\rho = 1.0$) on these surrogate subspaces. As shown in Table 9, all attacked checkpoints remain highly detectable with 100% Bit Accuracy and negligible perplexity degradation. This confirms that independent-data reconstruction neither recovers the owner's realized $\mathrm{Span}(U^\star)$ nor enables targeted removal without the exact private keys.

Table 9: Robustness against public-information subspace-reconstruction attack ($\rho = 1.0$). Values are mean $\pm$ standard deviation over three runs. The surrogate subspace estimated from independent data fails to align with the true watermark-bearing directions.

Model	Pre-attack S	Post-attack S	Threshold	Bit Acc	Pass	Relative ΔPPL
LLaMA-2-7B	5.766 ± 1.752	5.786 ± 1.712	1.954	100%	3/3	$+0.024$
Qwen2.5-7B	4.948 ± 0.334	4.958 ± 0.343	1.024	100%	3/3	$+0.004$

G SUPPLEMENTARY FIGURES

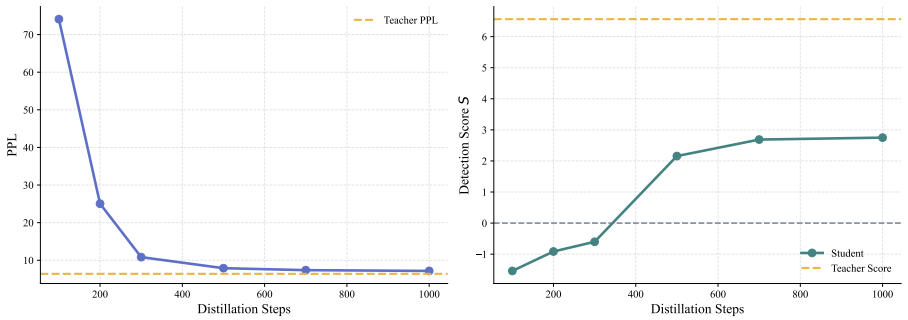


Figure 4: Step-wise behavior under homogeneous unconstrained distillation. This diagnostic single-seed sweep visualizes the coupling between student utility recovery and residual watermark signal.

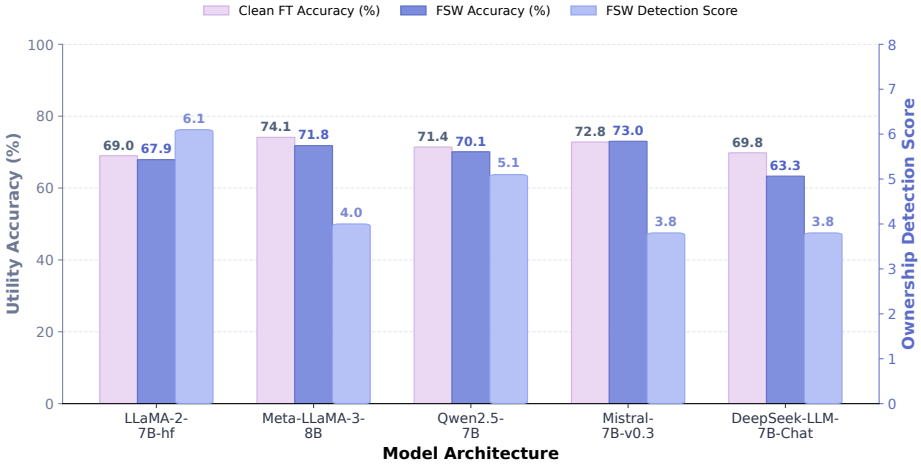


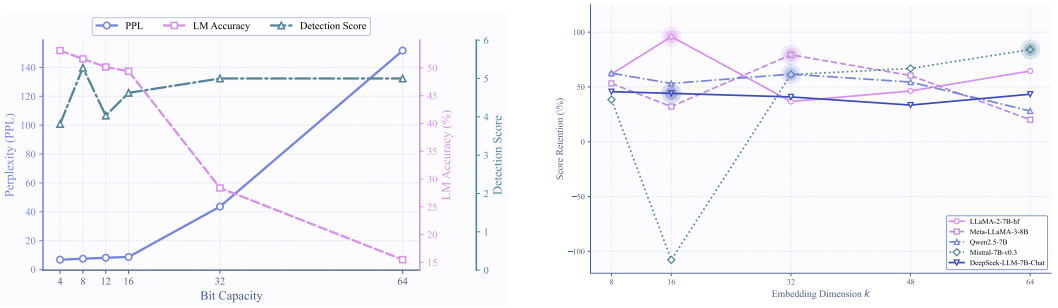
Figure 5: Functional preservation and ownership detectability across representative LLM backbones.

Cross-Backbone Utility and Detectability. Figure 5 provides a visual summary of functional preservation and ownership detectability across representative LLM backbones. Consistent with the main results, FSW maintains competitive downstream utility on most backbones while producing clearly positive detection scores. The main exception is DeepSeek-LLM-7B-Chat, where the utility drop is more visible, indicating that the utility–detectability trade-off can be architecture-dependent. Overall, the figure supports that FSW is not tied to a single model family, while also highlighting the need to account for backbone-specific sensitivity.

Payload and Embedding-Dimension Trade-offs. Figure 6 analyzes how watermark payload size and embedding dimension affect the utility–detectability trade-off. Moderate payload sizes maintain controlled perplexity and positive detection scores, whereas overly large payloads can substantially degrade language-modeling utility even when the watermark remains detectable. The embedding dimension study further shows that the optimal subspace size is backbone-dependent: too small a subspace may not contain enough stable watermark carriers, while too large a subspace may include fragile or utility-sensitive directions. These results motivate selecting payload and subspace dimension according to a validation utility budget.

H SUPPLEMENTARY TABLES

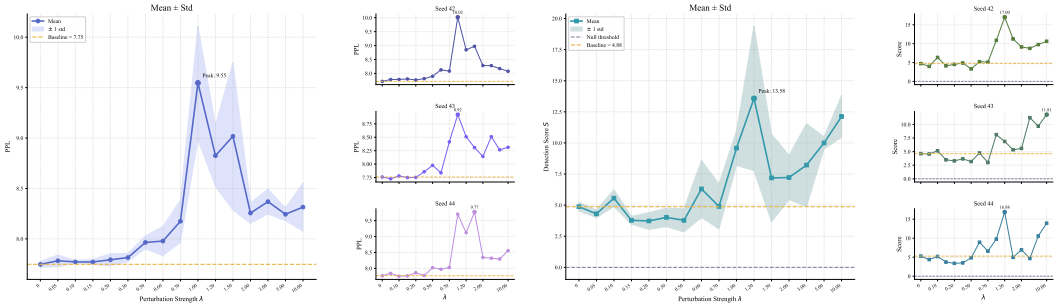
Full Robustness Results. Table 12 reports the full per-backbone robustness results under common model-side modifications. This table complements the compact robustness summary in the main text by showing detection scores, score retention, bit accuracy, perplexity, and perplexity changes



(a) Bit capacity vs. utility and detectability.

(b) Robustness–capacity trade-off under different embedding dimensions.

Figure 6: Trade-off analyses of FSW under varying bit capacity and embedding dimension.



(a) PPL under perturbation.

(b) Detection score under perturbation.

Figure 7: Untargeted representation perturbation stress test. Without instance-level verification secrets, perturbations weaken detection at moderate budgets and increase perplexity at larger budgets.

Table 10: Statistical verification of the 500-step homogeneous distillation student, showing detectable watermark signal across significance levels.

Significance α	Threshold T_α	Score S	Margin $S - T_\alpha$	Detected
10^{-2}	0.956 ± 0.000	2.542 ± 1.010	1.586 ± 1.010	Yes
10^{-3}	1.270 ± 0.000	2.542 ± 1.010	1.272 ± 1.010	Yes
10^{-4}	1.529 ± 0.000	2.542 ± 1.010	1.013 ± 1.010	Yes
10^{-6}	1.954 ± 0.000	2.542 ± 1.010	0.588 ± 1.010	Yes
10^{-8}	2.307 ± 0.000	2.542 ± 1.010	0.235 ± 1.010	Yes

Table 11: Full spectral position selection ablation under backbone distillation on LLaMA-2-7B-hf.

Selection	Pre PPL↓	Pre S↑	Post PPL↓	Post S↑	ΔS	Retention↑	Pre Bit Acc↑	Post Bit Acc↑
Top-k	6.5974	3.2500	6.6026	2.6250	-0.6250	80.77%	0.875	0.750
Adaptive	**5.9133**	**6.0938**	**5.9139**	**5.8438**	**-0.2500**	**95.90%**	**1.000**	**0.875**
Mid-50%	6.5938	4.8750	6.9178	3.6094	-1.2656	74.04%	1.000	1.000
Low-75%	6.0476	4.8438	6.4702	3.1719	-1.6719	65.48%	1.000	1.000
Last-k	6.5603	3.1406	6.4997	2.4531	-0.6875	78.11%	0.875	0.625

for each backbone–attack pair. Overall, the watermark signal remains recoverable across fine-tuning, noise perturbation, pruning, quantization, and backbone-preserving distillation, although the strength of retention varies across architectures and attack types.

Table 12: Robustness under model-side attacks. Ownership signals remain recoverable across diverse LLM backbones after fine-tuning, pruning, quantization, and backbone distillation.

Model	Attack	Watermark Reliability			Model Quality	
		Det. Score↑	Ret.↑	Bit Acc↑	PPL↓	ΔPPL↓
LLaMA-2-7B-hf	**Baseline**	**6.0938**	**100.00%**	**100.00%**	**5.91**	**+0.00**
	Backbone Distillation	5.8438	95.90%	87.50%	5.91	+0.00
	LoRA FT	6.7812	111.28%	87.50%	5.58	-0.33
	Noise	5.9062	96.92%	100.00%	5.92	+0.01
	Pruning	6.7188	110.26%	87.50%	6.13	+0.22
	Quantization	6.5625	107.69%	100.00%	6.43	+0.51
Meta-LLaMA-3-8B	**Baseline**	**4.0000**	**100.00%**	**100.00%**	**6.20**	**+0.00**
	Backbone Distillation	3.1719	79.30%	62.50%	6.12	-0.08
	LoRA FT	4.2188	105.47%	87.50%	6.10	-0.10
	Noise	4.1250	103.12%	100.00%	6.21	+0.01
	Pruning	3.2812	82.03%	87.50%	6.27	+0.07
	Quantization	4.2500	106.25%	87.50%	6.65	+0.45
Qwen2.5-7B	**Baseline**	**5.0938**	**100.00%**	**100.00%**	**7.24**	**+0.00**
	Backbone Distillation	3.1406	61.66%	75.00%	7.66	+0.42
	LoRA FT	3.4688	68.10%	87.50%	6.75	-0.49
	Noise	4.6875	92.02%	100.00%	7.26	+0.02
	Pruning	4.8438	95.09%	100.00%	7.32	+0.08
	Quantization	5.2500	103.07%	100.00%	8.22	+0.98
Mistral-7B-v0.3	**Baseline**	**3.8438**	**100.00%**	**100.00%**	**17.25**	**+0.00**
	Backbone Distillation	3.2344	84.15%	75.00%	6.59	-10.66
	LoRA FT	3.7812	98.37%	100.00%	16.92	-0.33
	Noise	3.7656	97.97%	100.00%	17.19	-0.05
	Pruning	3.3281	86.59%	100.00%	17.58	+0.34
	Quantization	3.7969	98.78%	100.00%	18.99	+1.75
DeepSeek-LLM-7B-Chat	**Baseline**	**3.7500**	**100.00%**	**100.00%**	**77.31**	**+0.00**
	Backbone Distillation	1.6562	44.17%	100.00%	61.39	-15.92
	LoRA FT	0.6602	17.60%	100.00%	62.31	-15.00
	Noise	3.7344	99.58%	100.00%	77.28	-0.03
	Pruning	3.4844	92.92%	100.00%	81.79	+4.48
	Quantization	3.7188	99.17%	100.00%	92.76	+15.45

Full Comparison with SEAL. Table 13 reports the complete comparison between FSW and SEAL under aligned data, utility evaluation, and attack settings. Both methods are evaluated using their intended payloads and native verification rules. Because their detection statistics are method-specific, we compare perplexity, configuration-specific bit accuracy, and native verification outcomes rather than raw detection scores.

ECC Analysis. We further evaluate whether lightweight error-correcting codes can improve the reliability of multi-bit message recovery. Table 15 reports bit-level and message-level decoding accuracy on LLaMA-2-7B-hf with and without Hamming (7,4) correction. ECC is most useful in borderline payload regimes: at 8 bits, the raw watermark recovery contains residual bit errors, while Hamming correction restores both bit accuracy and message accuracy to 100%. At other payload sizes, raw recovery is already near-perfect or the coding overhead is less beneficial. These results suggest that ECC can serve as a lightweight reliability enhancement rather than a substitute for stable watermark embedding.

Dataset Generalization. Table 16 reports the generalization behavior of FSW across different fine-tuning corpora on Meta-LLaMA-3-8B. Although shifting from WikiText-2 to C4 increases perplexity, FSW still preserves strong detectability, including 100% bit accuracy and near-perfect AUC.

Table 13: Complete comparison between FSW and SEAL. Results are mean ± standard deviation over three runs. FSW and SEAL use their intended 8-bit and 64-bit payloads, respectively. Bit accuracy is reported as a fraction. Mean Det. indicates whether the mean FSW detection score exceeds its corresponding threshold, while Native Passes reports the number of SEAL runs satisfying its native verification rule.

Model	Attack	FSW (8-bit)		SEAL (64-bit)	
		PPL / Bit Acc.	Mean Det.	PPL / Bit Acc.	Native Passes
LLaMA-2-7B-hf	No Attack	5.961 ± 0.071 / 1.000 ± 0.000	Yes	14.378 ± 0.261 / 1.000 ± 0.000	3/3
	LoRA	5.569 ± 0.010 / 0.875 ± 0.125	Yes	9.061 ± 0.028 / 1.000 ± 0.000	3/3
	Quantization	6.426 ± 0.014 / 0.875 ± 0.217	Yes	15.751 ± 0.278 / 1.000 ± 0.000	3/3
	Noise	5.967 ± 0.065 / 1.000 ± 0.000	Yes	14.393 ± 0.265 / 1.000 ± 0.000	3/3
	Pruning	6.094 ± 0.033 / 0.917 ± 0.072	Yes	15.814 ± 0.327 / 1.000 ± 0.000	3/3
	BPD	5.982 ± 0.063 / 0.875 ± 0.125	Yes	9.825 ± 0.112 / 0.531 ± 0.072	0/3
	Hom.-500	7.995 ± 0.338 / 0.625 ± 0.000	Yes	11.524 ± 0.186 / 0.557 ± 0.048	0/3
Qwen2.5-7B	No Attack	7.201 ± 0.132 / 1.000 ± 0.000	Yes	13.408 ± 0.055 / 1.000 ± 0.000	3/3
	LoRA	6.754 ± 0.019 / 0.833 ± 0.072	Yes	10.378 ± 0.015 / 1.000 ± 0.000	3/3
	Quantization	8.158 ± 0.150 / 1.000 ± 0.000	Yes	14.756 ± 0.027 / 1.000 ± 0.000	3/3
	Noise	7.204 ± 0.131 / 1.000 ± 0.000	Yes	13.402 ± 0.055 / 1.000 ± 0.000	3/3
	Pruning	7.276 ± 0.133 / 1.000 ± 0.000	Yes	13.551 ± 0.045 / 1.000 ± 0.000	3/3
	BPD	7.664 ± 0.110 / 0.750 ± 0.000	Yes	11.984 ± 0.036 / 0.516 ± 0.031	0/3
	Hom.-500	8.145 ± 0.025 / 0.500 ± 0.000	No	11.884 ± 0.028 / 0.542 ± 0.024	0/3

Table 14: Step-wise results of homogeneous unconstrained distillation from watermarked teachers to clean-base students across backbones. Values are mean ± standard deviation over runs.

Backbone	Model	Steps	Runs	PPL↓	HellaSwag↑	ARC-E↑	Score S ↑
LLaMA-2-7B (top-k-logits=200)	FSW Teacher	–	3	5.913 ± 0.000	0.648 ± 0.000	0.710 ± 0.000	6.094 ± 0.000
	Clean-base Student	500	3	7.995 ± 0.338	0.477 ± 0.003	0.757 ± 0.003	2.542 ± 1.010
	Clean-base Student	1000	3	7.098 ± 0.123	0.483 ± 0.003	0.750 ± 0.010	2.417 ± 0.048
	Clean-base Student	1500	3	6.767 ± 0.034	0.478 ± 0.003	0.753 ± 0.003	1.706 ± 0.711
Qwen-2.5-7B (top-k-logits=1000)	FSW Teacher	–	3	7.244 ± 0.000	0.668 ± 0.000	0.734 ± 0.000	5.938 ± 0.000
	Clean-base Student	500	3	8.145 ± 0.025	0.490 ± 0.005	0.805 ± 0.010	-1.813 ± 0.299
	Clean-base Student	1000	3	7.848 ± 0.012	0.490 ± 0.000	0.800 ± 0.009	1.040 ± 0.412
	Clean-base Student	1500	3	7.838 ± 0.017	0.487 ± 0.003	0.810 ± 0.005	1.656 ± 0.521

Table 15: ECC improves message decoding accuracy for LLaMA-2-7B-hf, correcting residual errors at 8 bits.

Bits	No ECC (%)		Hamming (7,4) (%)	
	Bit Acc ↑	Msg Acc ↑	Bit Acc ↑	Msg Acc ↑
4	100.0	100.0	0.0	0.0
8	87.5	87.5	**100.0**	**100.0**
12	100.0	100.0	100.0	100.0
16	100.0	100.0	93.8	100.0
32	100.0	100.0	100.0	100.0
64	100.0	100.0	100.0	100.0

This suggests that the identified functional backbone is not overly tied to a specific fine-tuning corpus and remains effective under distributional variation.

Table 16: Dataset generalization on Meta-LLaMA-3-8B. FSW maintains competitive utility and strong detectability across different fine-tuning corpora.

Dataset	Method	Functional Preservation				Detectability		
		PPL↓	ΔPPL↓	HellaSwag↑	ARC-E↑	Det. Score↑	Bit Acc↑	AUC↑
WikiText-2	Clean FT	5.83	0.00	68.60	79.60	-0.03	–	–
	FSW (Ours)	6.20	+0.37	67.80	75.80	**4.00**	**100%**	0.899
C4	Clean FT	8.69	0.00	79.84	77.74	-0.01	–	–
	FSW (Ours)	9.59	+0.90	77.38	69.28	**4.88**	**100%**	1.000

Significance Test under Different Detection Thresholds. Table 17 summarizes the statistical verification results under different significance levels α. For both Mistral-7B and DeepSeek-7B, the detection score remains above the critical threshold over a broad range of settings, indicating reliable ownership verification. In particular, DeepSeek-7B preserves a large positive margin even under very stringent significance constraints, further supporting the statistical reliability of FSW.

Table 17: Significance test results under different levels of α. FSW maintains a positive detection margin under stringent significance constraints across different backbones.

Model	Significance Level (α)	Threshold T_α	Detection Score S	Margin ($S - T_\alpha$)	Detected
LLaMA-2-7B	1e−02	0.9564	6.0938	5.1374	✓
	1e−03	1.2704	6.0938	4.8234	✓
	1e−04	1.5289	6.0938	4.5649	✓
	1e−06	1.9541	6.0938	4.1397	✓
	1e−08	2.3071	6.0938	3.7867	✓
Mistral-7B	1e−02	0.4391	3.8438	3.4047	✓
	1e−03	0.5833	3.8438	3.2605	✓
	1e−04	0.7020	3.8438	3.1418	✓
	1e−06	0.8972	3.8438	2.9466	✓
	1e−08	1.0593	3.8438	2.7845	✓
DeepSeek-7B	1e−02	0.1167	3.7500	3.6333	✓
	1e−03	0.1550	3.7500	3.5950	✓
	1e−04	0.1865	3.7500	3.5635	✓
	1e−06	0.2384	3.7500	3.5116	✓
	1e−08	0.2815	3.7500	3.4685	✓

Baseline Comparison under Common Function-Preserving Modifications. Table 18 compares FSW with representative watermarking baselines before and after distillation. While several baselines exhibit substantial degradation in either utility or robustness, FSW maintains a more favorable balance between perplexity and downstream performance after attack. These results support the claim that functional-subspace-based embedding is more resilient under backbone-preserving distillation than conventional alternatives.

Comprehensive Scan of Embedding Dimension k. Table 19 presents a detailed scan of the embedding dimension k across all evaluated backbones. The results show that the optimal k varies substantially by model, reflecting differences in representation geometry and robustness bottlenecks. Smaller values may fail to capture enough stable carrier directions, whereas overly large values may overlap with sensitive semantic dimensions and reduce robustness after attack.

Optimal Embedding Dimension Selection. Table 20 summarizes the final choice of embedding dimension for each backbone model. The selected k maximizes post-attack score retention while maintaining acceptable utility, and therefore serves as the default configuration in the main experiments. This table provides a compact summary of the broader scan reported in Table 19.

Table 18: **Baseline Comparison under Common Function-Preserving Modifications.** We report perplexity (PPL) and downstream accuracies (HellaSwag and ARC-Challenge) before and after distillation for multiple watermarking baselines.

Model	Variant	Original (Pre-Attack)			Attacked (Post-Attack)		
		PPL↓	Hella↑	ARC↑	PPL↓	Hella↑	ARC↑
Llama-2-7B	Clean FT (No WM)	5.30	66.40%	71.60%	—	—	—
	FSW	5.91	64.80%	71.00%	5.91	65.00%	71.20%
	EmMark	416.15	65.80%	71.40%	5.73	65.80%	72.60%
	Weight Quant.	20.88	66.60%	71.60%	5.78	68.00%	72.40%
	Naive Top-k	7.30	66.80%	72.20%	6.75	66.60%	71.60%
Llama-3-8B	Clean FT (No WM)	5.83	68.60%	79.60%	—	—	—
	FSW	6.20	67.80%	75.80%	6.12	68.00%	76.00%
	EmMark	23.24	68.60%	77.40%	6.09	69.00%	77.80%
	Weight Quant.	9.77	69.60%	76.00%	6.03	69.80%	77.40%
	Naive Top-k	7.35	69.80%	74.20%	6.64	70.00%	76.40%
Qwen2.5-7B	Clean FT (No WM)	6.35	67.00%	75.80%	—	—	—
	FSW	7.24	66.80%	73.40%	7.66	66.80%	73.20%
	EmMark	40.62	67.00%	72.00%	6.68	67.20%	78.60%
	Weight Quant.	24.02	66.80%	78.80%	6.66	66.80%	78.80%
	Naive Top-k	7.94	67.00%	77.20%	7.79	67.00%	77.00%
Mistral-7B	Clean FT (No WM)	16.68	68.60%	77.00%	—	—	—
	FSW	17.25	69.00%	77.00%	6.59	68.60%	76.40%
	EmMark	29.20	71.80%	78.40%	6.24	70.60%	77.20%
	Weight Quant.	27.49	71.40%	79.60%	6.23	69.20%	76.60%
	Naive Top-k	21.24	70.40%	76.40%	6.93	67.20%	73.20%
DeepSeek-7B	Clean FT (No WM)	59.52	67.40%	72.20%	—	—	—
	FSW	77.31	60.40%	66.20%	61.39	60.60%	66.60%
	EmMark	152.36	69.60%	70.60%	7.95	70.20%	72.40%
	Weight Quant.	129.89	71.00%	70.20%	8.23	70.60%	71.20%
	Naive Top-k	87.12	67.40%	70.40%	67.00	67.00%	70.80%

Table 19: Comprehensive Scan of Embedding Dimension k across Different LLM Backbones. *Pre* and *post* denote metrics before and after the robustness attack (e.g., distillation or quantization). Ret.(%) indicates the score retention rate. **Bold** k values represent the optimal dimensions selected for each model.

Model	k	Original (Pre-Attack)			Attacked (Post-Attack)			Ret. (%)
		PPL↓	Score↑	Acc↑	PPL↓	Score↑	Acc↑	
Llama-2-7B	8	5.9661	5.0625	100.0%	6.0652	3.1250	50.0%	61.73
	16	5.9133	6.0938	100.0%	5.9139	5.8438	87.5%	**95.90**
	32	6.0213	4.9375	100.0%	6.0903	1.8125	62.5%	36.71
	48	5.9709	5.9375	87.5%	6.0322	2.7500	62.5%	46.32
	64	5.9611	5.1563	87.5%	6.2732	3.3281	87.5%	64.55
Llama-3-8B	8	6.1810	5.4063	100.0%	6.0825	2.8750	87.5%	53.18
	16	6.1936	5.1875	87.5%	6.0938	1.6641	75.0%	32.08
	32	6.2016	4.0000	100.0%	6.1220	3.1719	62.5%	**79.30**
	48	6.1880	4.5000	87.5%	6.0819	2.7188	62.5%	60.42
	64	6.1874	5.2500	100.0%	6.1162	1.0547	75.0%	20.09
Qwen2.5-7B	8	7.4219	5.5938	100.0%	7.9128	3.5000	62.5%	62.57
	16	7.1660	4.4375	100.0%	7.6541	2.3594	87.5%	53.17
	32	7.2437	5.0938	100.0%	7.6610	3.1406	75.0%	**61.66**
	48	7.4752	4.9063	100.0%	7.9819	2.6719	62.5%	54.46
	64	7.1649	5.3438	100.0%	7.8791	1.5000	62.5%	28.07
Mistral-7B	8	17.4032	4.1250	100.0%	6.5614	1.5859	62.5%	38.45
	16	17.2952	4.0313	100.0%	6.5578	-4.3438	25.0%	-107.75
	32	17.2976	5.2500	100.0%	6.5754	3.2188	87.5%	61.31
	48	17.3229	5.0000	100.0%	6.7003	3.3438	75.0%	66.88
	64	17.2461	3.8438	100.0%	6.5854	3.2344	75.0%	**84.15**
DeepSeek-7B	8	76.7636	3.7969	100.0%	58.3845	1.7344	100.0%	45.68
	16	77.3068	3.7500	100.0%	61.3864	1.6563	100.0%	**44.17**
	32	75.3895	3.8594	100.0%	68.7011	1.5781	100.0%	40.89
	48	80.5563	4.2500	100.0%	66.9643	1.4219	100.0%	33.46
	64	80.3019	4.1250	100.0%	71.6761	1.7891	100.0%	43.37

Table 20: Optimal embedding dimension k selection across backbone models. The optimal k is determined by maximizing the retention rate of detection scores after robustness attacks, ensuring an effective trade-off between watermark capacity and structural stability.

Model	Candidates	Opt. k	Pre-Attack (Original)			Post-Attack (Robustness)		
			PPL↓	Score↑	Acc↑	PPL↓	Score↑	Acc↑
LLaMA-2-7B-hf	$\{8, 16, 32, \dots\}$	**16**	5.91	6.09	100.0%	5.91	**5.84**	87.5%
LLaMA-3-8B	$\{8, 16, 32, \dots\}$	**32**	6.20	4.00	100.0%	6.12	**3.17**	62.5%
Qwen2.5-7B	$\{8, 16, 32, \dots\}$	**32**	7.24	5.09	100.0%	7.66	**3.14**	75.0%
Mistral-7B-v0.3	$\{8, 16, 32, \dots\}$	**64**	17.25	3.84	100.0%	6.59	**3.23**	75.0%
DeepSeek-7B-Chat	$\{8, 16, 32, \dots\}$	**16**	77.31	3.75	100.0%	61.39	**1.66**	100.0%

Boundary Stress-Test Details. We report the complete numerical results for the boundary analyses discussed in the main text. These experiments are intended to characterize robustness limits rather than to expand the formal threat model beyond function-preserving modifications.

Table 21: Full multi-seed representation perturbation results. Values are mean ± standard deviation over seeds 42, 43, and 44.

λ	PPL↓	Det. Score S ↑	Bit Acc↑
0.00	7.7476 ± 0.0298	4.8750 ± 0.3549	0.8750 ± 0.1250
0.05	7.7828 ± 0.0579	4.3021 ± 0.3035	0.7500 ± 0.0000
0.10	7.7717 ± 0.0154	5.5625 ± 0.6768	0.8750 ± 0.1250
0.15	7.7712 ± 0.0264	3.7813 ± 0.3171	0.7500 ± 0.0000
0.20	7.7930 ± 0.0601	3.7292 ± 0.6692	0.7500 ± 0.0000
0.30	7.8146 ± 0.0404	4.0156 ± 0.7492	0.7917 ± 0.0722
0.50	7.9647 ± 0.0630	3.7761 ± 0.9274	0.7917 ± 0.0722
0.60	7.9786 ± 0.1447	6.3021 ± 2.2943	0.8333 ± 0.0722
0.70	8.1751 ± 0.2096	4.8959 ± 1.8051	0.7917 ± 0.1443
1.00	9.5468 ± 0.5637	9.5833 ± 1.3826	0.7083 ± 0.0722
1.20	8.8256 ± 0.3047	13.5833 ± 5.8099	0.6667 ± 0.0722
1.50	9.0175 ± 0.7317	7.1875 ± 3.5232	0.6250 ± 0.0000
2.00	8.2561 ± 0.1024	7.2188 ± 1.7823	0.6250 ± 0.0000
3.00	8.3693 ± 0.1238	8.2187 ± 3.3288	0.5833 ± 0.0722
5.00	8.2443 ± 0.0639	10.0000 ± 0.4881	0.6250 ± 0.1250
10.00	8.3152 ± 0.2385	12.1250 ± 1.6782	0.7083 ± 0.0722

Table 22: Training overhead of FSW compared to standard fine-tuning. While embedding introduces additional memory requirements due to dual-model consistency constraints, the overall time overhead remains manageable for 7B-scale models.

Training Phase (Qwen2.5-7B)	Time (s)	Steps	GPU Mem (MB)	CPU Mem (MB)
Standard FT (Baseline)	1101.55	7500	29820.63	6111.39
Phase 1: Fisher Estimation	76.61	250	29306.54	20815.56
Phase 2: Inv. Estimation	11.66	250	29287.93	20815.56
Phase 3: Subspace Construction	3.02	1	29268.78	20815.56
Phase 4: Watermark Embedding	1764.38	7500	56410.34	20815.56
Total FSW	**1855.68**	**8001**	**56410.34**	**20815.56**

Training Overhead Analysis. Table 22 reports the computational overhead of FSW relative to standard fine-tuning on Qwen2.5-7B. Most of the additional cost is concentrated in the watermark embedding phase, where the consistency constraint requires maintaining both the current model and the frozen reference representation. Despite this added memory and time cost, the overall overhead remains manageable for 7B-scale models.

Model Specifications and Watermark Layer Selection. Table 7 summarizes the architectural specifications of all evaluated models together with the selected watermark layer ℓ. In all cases, the watermark is injected at approximately the middle layer, which empirically provides a favorable balance between representational stability and semantic expressiveness. This configuration is used consistently throughout the experiments unless otherwise stated.

Table 23: Per-seed representation perturbation results. The attack uses LoRA rank 8, 100 optimization steps, C4 attack data, and WikiText-2 test perplexity evaluation.

Seed	λ	PPL↓	HellaSwag↑	ARC-E↑	Det. Score S ↑	Bit Acc↑
42	0.00	7.7137	0.485	0.760	4.7188	0.875
42	0.05	7.7809	0.480	0.765	3.9688	0.750
42	0.10	7.7826	0.475	0.755	6.3438	1.000
42	0.15	7.8003	0.475	0.755	4.1250	0.750
42	0.20	7.7658	0.475	0.755	4.5000	0.750
42	0.30	7.8106	0.480	0.755	4.8750	0.750
42	0.50	7.8967	0.480	0.755	3.3125	0.750
42	0.60	8.1265	0.485	0.755	5.2188	0.875
42	0.70	8.0840	0.480	0.750	5.0938	0.875
42	1.00	10.0178	0.460	0.715	10.8750	0.750
42	1.20	8.8464	0.480	0.720	17.0000	0.750
42	1.50	8.9752	0.470	0.720	11.2500	0.625
42	2.00	8.2828	0.475	0.745	9.1250	0.625
42	3.00	8.2778	0.475	0.740	8.7500	0.625
42	5.00	8.1723	0.470	0.745	9.7500	0.750
42	10.00	8.0777	0.470	0.750	10.6250	0.750
43	0.00	7.7591	0.485	0.755	4.6250	0.750
43	0.05	7.7258	0.475	0.760	4.5625	0.750
43	0.10	7.7783	0.485	0.755	5.1562	0.750
43	0.15	7.7490	0.485	0.765	3.5000	0.750
43	0.20	7.7512	0.480	0.760	3.2969	0.750
43	0.30	7.8569	0.485	0.755	3.6719	0.750
43	0.50	7.9764	0.485	0.755	3.1719	0.750
43	0.60	7.8374	0.480	0.735	4.7500	0.750
43	0.70	8.4148	0.485	0.740	3.0000	0.625
43	1.00	8.9222	0.470	0.720	8.1250	0.750
43	1.20	8.5111	0.475	0.740	6.8750	0.625
43	1.50	8.3079	0.470	0.725	5.3438	0.625
43	2.00	8.1430	0.480	0.740	5.5938	0.625
43	3.00	8.5102	0.475	0.735	11.2500	0.625
43	5.00	8.2662	0.475	0.740	9.6875	0.625
43	10.00	8.3132	0.475	0.740	11.8125	0.750
44	0.00	7.7699	0.480	0.765	5.2812	1.000
44	0.05	7.8416	0.485	0.760	4.3750	0.750
44	0.10	7.7541	0.490	0.760	5.1875	0.875
44	0.15	7.7642	0.485	0.765	3.7188	0.750
44	0.20	7.8619	0.480	0.765	3.3906	0.750
44	0.30	7.7763	0.490	0.765	3.5000	0.875
44	0.50	8.0211	0.480	0.755	4.8438	0.875
44	0.60	7.9720	0.480	0.755	8.9375	0.875
44	0.70	8.0264	0.480	0.745	6.5938	0.875
44	1.00	9.7005	0.460	0.715	9.7500	0.625
44	1.20	9.1194	0.475	0.720	16.8750	0.625
44	1.50	9.7694	0.485	0.735	4.9688	0.625
44	2.00	8.3424	0.475	0.740	6.9375	0.625
44	3.00	8.3200	0.475	0.730	4.6562	0.500
44	5.00	8.2944	0.475	0.740	10.5625	0.500
44	10.00	8.5546	0.470	0.735	13.9375	0.625

Table 24: Extended single-seed diagnostic sweep and convergence analysis for homogeneous unconstrained distillation on Qwen2.5-7B (top-k-logits=1000). The early high-PPL regime indicates that watermark weakening and utility recovery are coupled. After validation loss stabilizes (around 1500-2000 steps), the watermark signal is recovered and successfully passes verification.

Steps	Val. Loss	PPL↓	Det. Score S ↑	Margin ↑	Bit Acc↑	Detected
100	0.4497	74.1000	-1.5391	-2.5634	0.500	No
300	0.3769	10.8370	-0.6016	-1.6259	0.375	No
500	0.3300	7.8939	2.1563	1.1320	0.625	Yes
1000	0.3199	7.1682	2.7500	1.7257	0.625	Yes
1500	0.3150	7.8262	1.2734	0.2492	0.625	Yes
1750	0.3166	7.8622	2.1250	1.1007	0.625	Yes
2000	0.3175	7.8525	1.8828	0.8585	0.750	Yes